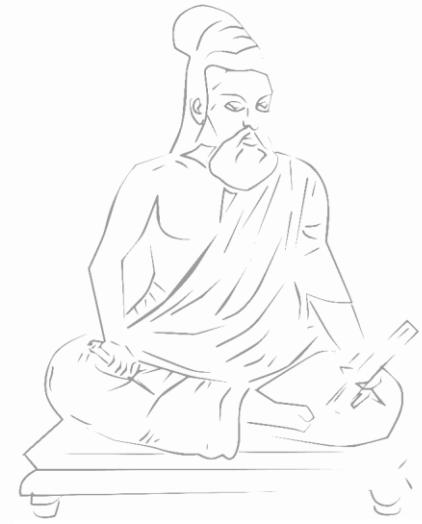


Azure Developer
Community

Tamil Nadu



What we do?



Meetups



Network of Beginners,
Professionals &
Experts



Azure Cert Study Jam



Tech-talks / workshops
In Colleges

RAG Is Not About Vector Databases - It's About Retrieval Thinking

Why Most GenAI Proof-of-
Concepts Fail in Production



Azure Developer
Community

Tamil Nadu





Self intro



If vector databases solved RAG...
why do most RAG apps still fail in
production?



Why LLMs Alone Are Not Enough

Trained on general internet data

No access to your private documents

Hallucinates when uncertain



What is RAG?



RAG = Retrieval + Generation



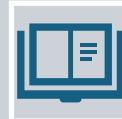
Retrieve relevant information



Generate answer using that information



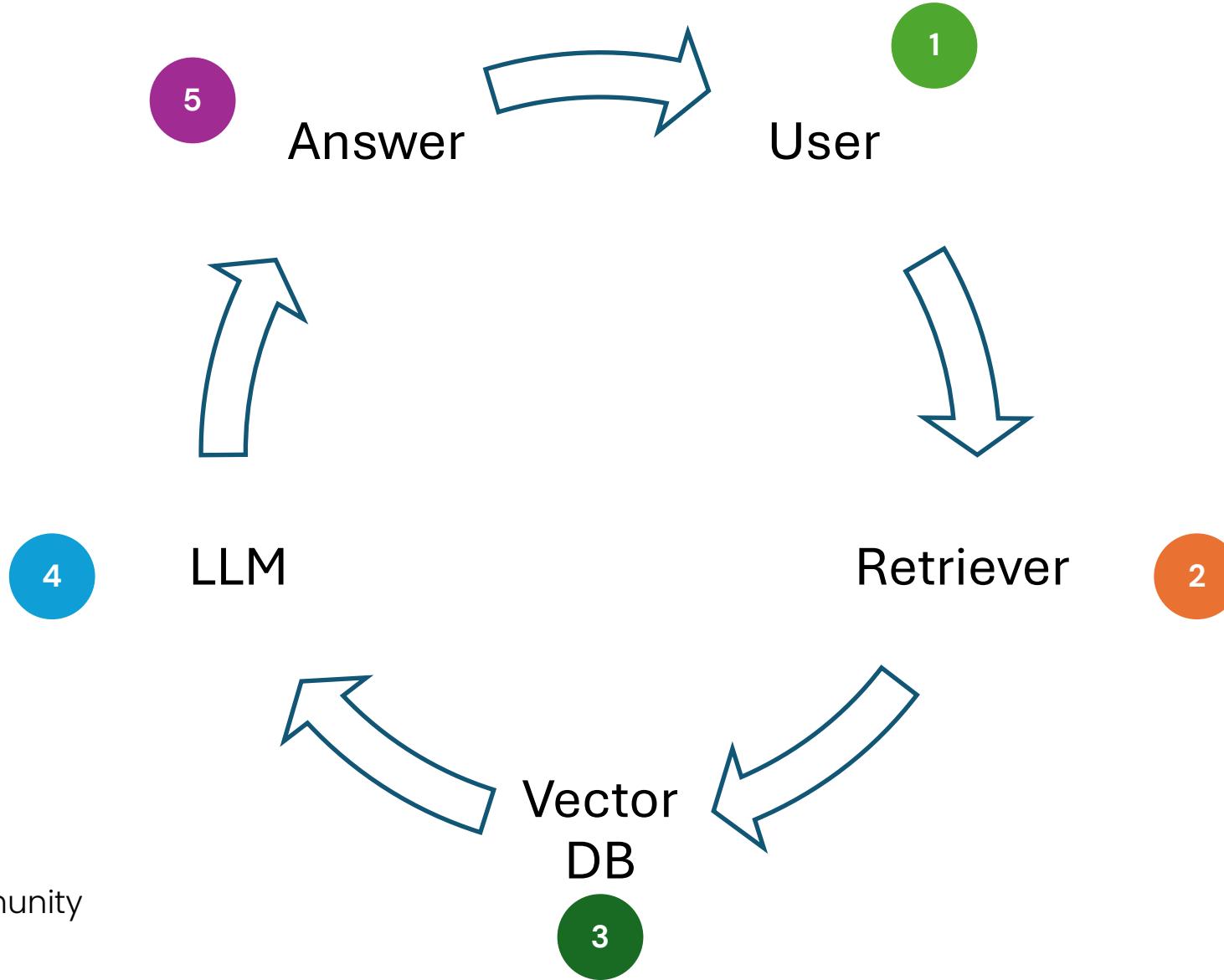
Analogy:



Open-book exam for the model



Typical RAG Architecture



The Problem



The Myth

#Azure Developer Community



“Just add a vector database”



“Better embeddings = better answers”



“More documents = better results”



Why PoCs Look Impressive



Small dataset



Known questions



Clean documents



Manual testing



Reality in Production

Messy documents

Unknown questions

Large datasets

Latency constraints



Retrieval Thinking



Retrieval Thinking

Before storing documents,
ask:

What kind of questions will users ask?

What granularity is needed?

What metadata matters?

What if answer does not exist?



Chunking Is a Design Decision

Bad:

- Large 2000-token chunks

Better:

- Logical sections
- Policy-based segmentation
- FAQ-style chunks



The Demo



After Basic RAG Failure



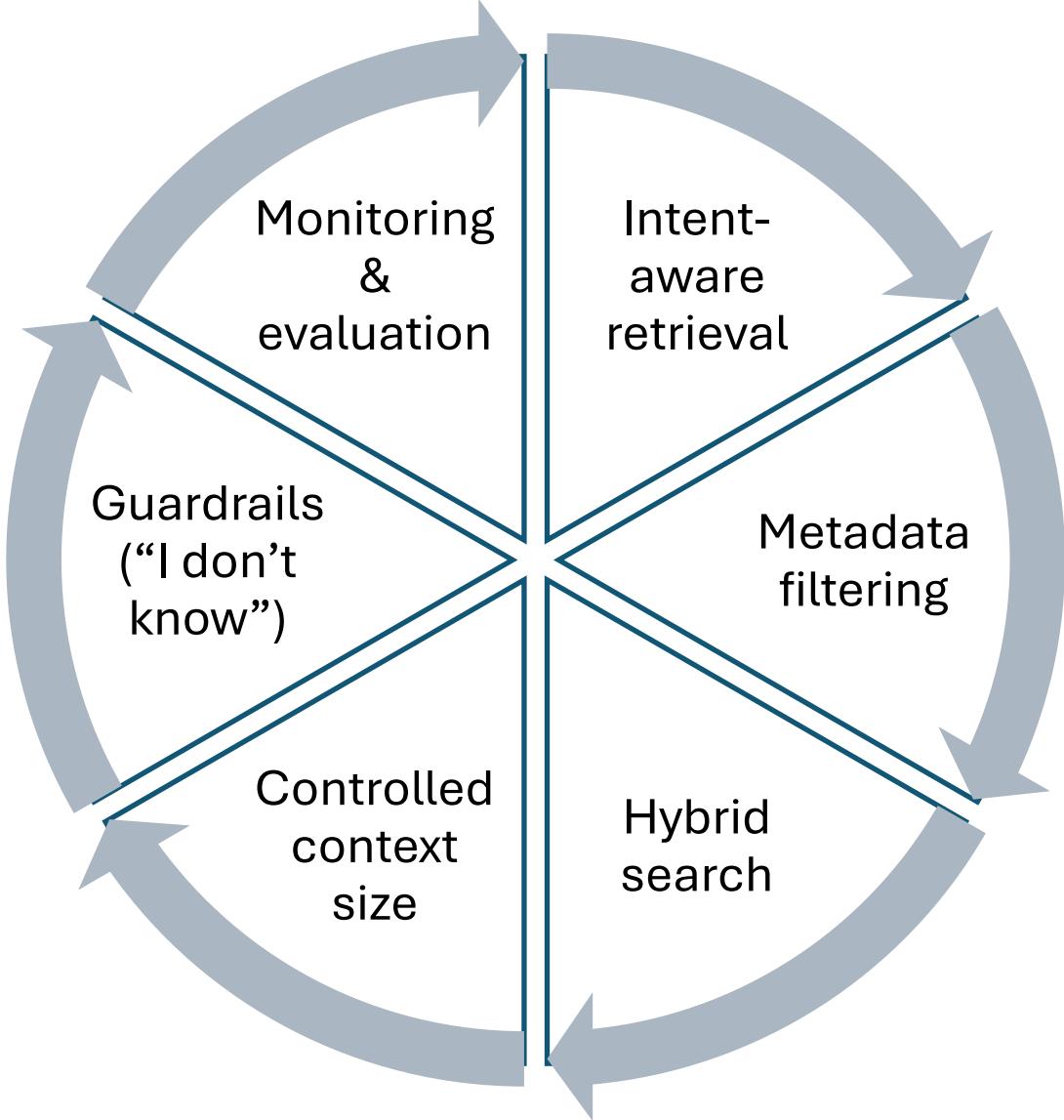
After Improved Retrieval



Production-Ready RAG



Production- Ready RAG Needs



RAG Patterns in Production



Hybrid RAG

Vector search + Keyword search
Better precision



Multi-Stage Retrieval

Retrieve broadly
Re-rank
Then generate



Agentic RAG

Planner decides which retriever to use
Tool-based retrieval



RAG + Evaluation

Automatic answer scoring
Feedback loops
Monitoring hallucinations



Taking RAG to Production



Latency constraints



Cost control (token usage)



Caching strategies



Embedding refresh pipelines



Document update handling



Access control (who can see what)

Final Takeaways

RAG is system design

Vector DB is just storage

Retrieval determines accuracy

Better context > Bigger model



Cohort

<https://www.learnwithsarvesh.com/>



LIVE SESSION

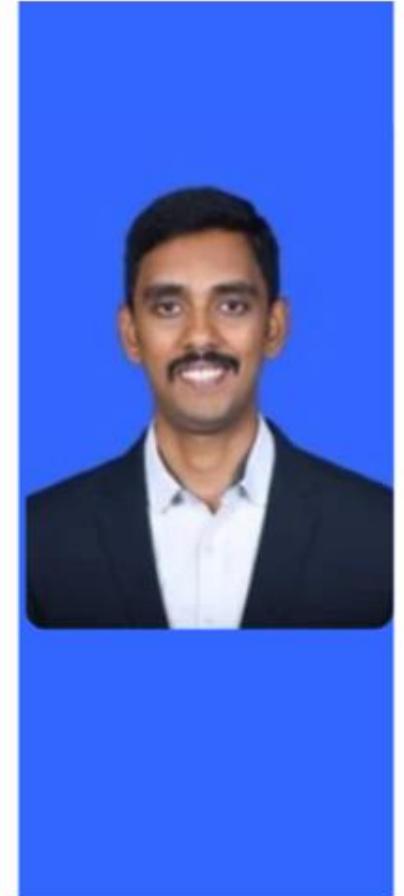
12-WEEK LANGCHAIN CRASH COURSE

Build Real-World AI Projects with Mentorship from Sarvesh

⌚ 24 hrs

WITH
SARVESH

🔗 Start Learning Today on [learnwithsarvesh.com](https://www.learnwithsarvesh.com)

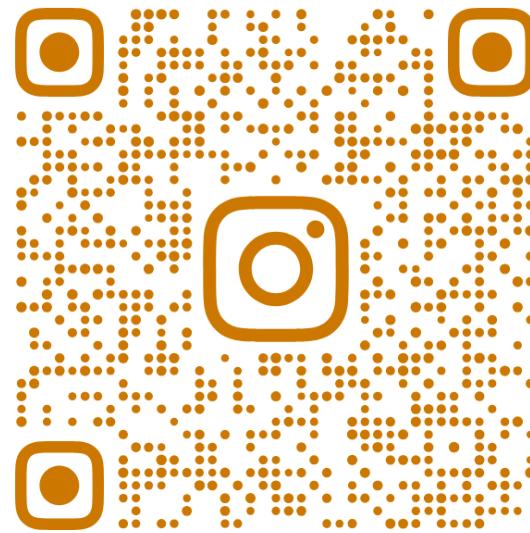


Open for Q&A



INGRAM[®]
MICRO

#Azure Developer Community



@WITHSARVESH.AI

Thank you!!

Sarveshwaran Rajagopal
Data Scientist and Trainer (AI Agents, RAG) | Empowered 7000+ Pro...



Thank You!

