TECHX

Asia's Largest

AI & Cloud

Conference 2024

15 - 16, November 2024

Chennai Trade Center, Chennai

Bhuvaneswari Subramani

Chief Cloud Evangelist @ Intuitive.Cloud

AWS Hero, AWS Ambassador

I am a technology leader with 24 years of IT experience, specializing in Cloud Modernization, DevOps, Cloud Alliance, and Cloud Financial Management.
I fervently drive Global Outreach programs. I am also a passionate blogger and an active speaker on Technology and Leadership in tech communities, international conferences, TEDx, and universities.

aws HEROES

TEDx Speaker

aws AMBASSADOR

Agentic RAG

A Self-Corrective Method for Implementing Retrieval-Augmented Generation

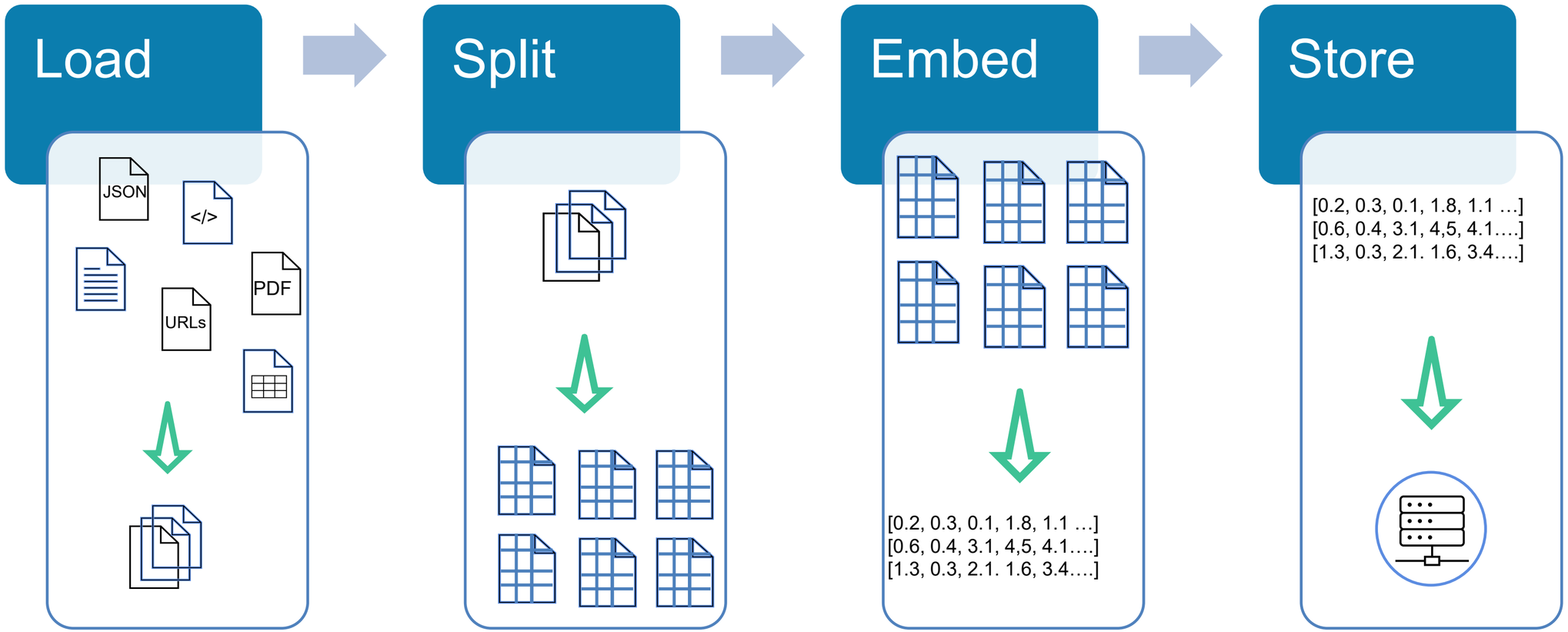
1. Traditional RAG System

Traditional RAG System

1. Data Processing and Indexing

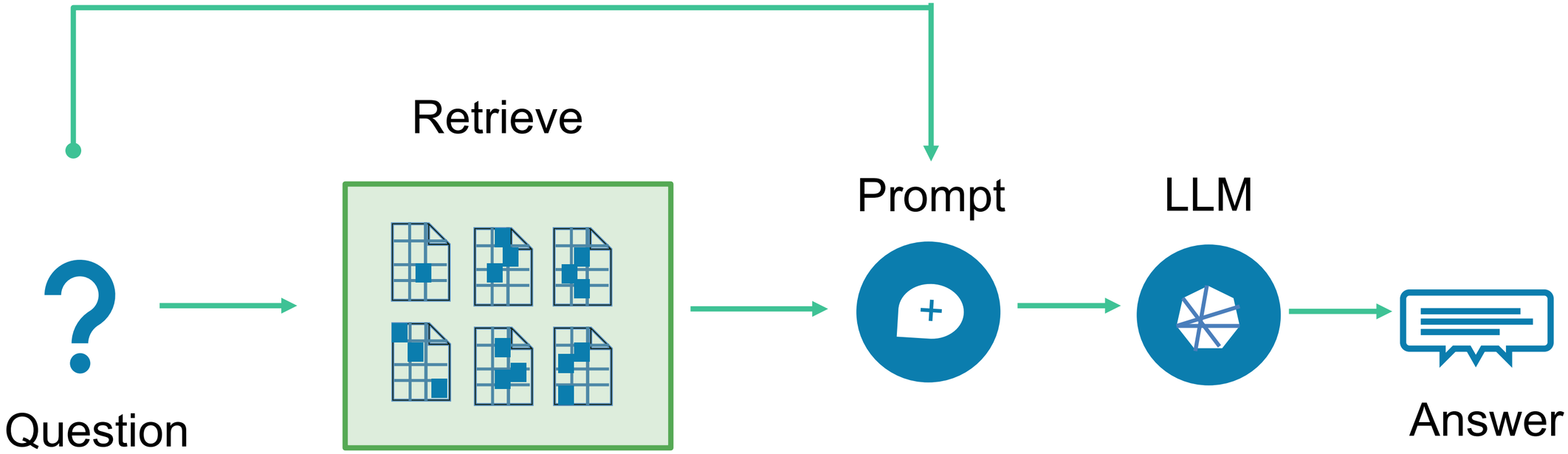
2. Retrieval and Response Generation

Data Processing & Indexing



https://www.analyticsvidhya.com/blog/2024/07/building-agentic-rag-systems-with-langgraph/

Retrieval and Response Generation



Limitations of Traditional RAG

Traditional RAG systems face several challenges, including:

- Lack of access to real-time data.
- System dependence on the quality of data in vector database.
- Ineffective retrieval strategies may result in irrelevant documents being used for responses.
- Large language models (LLMs) may produce hallucinations or fail to answer questions accurately.

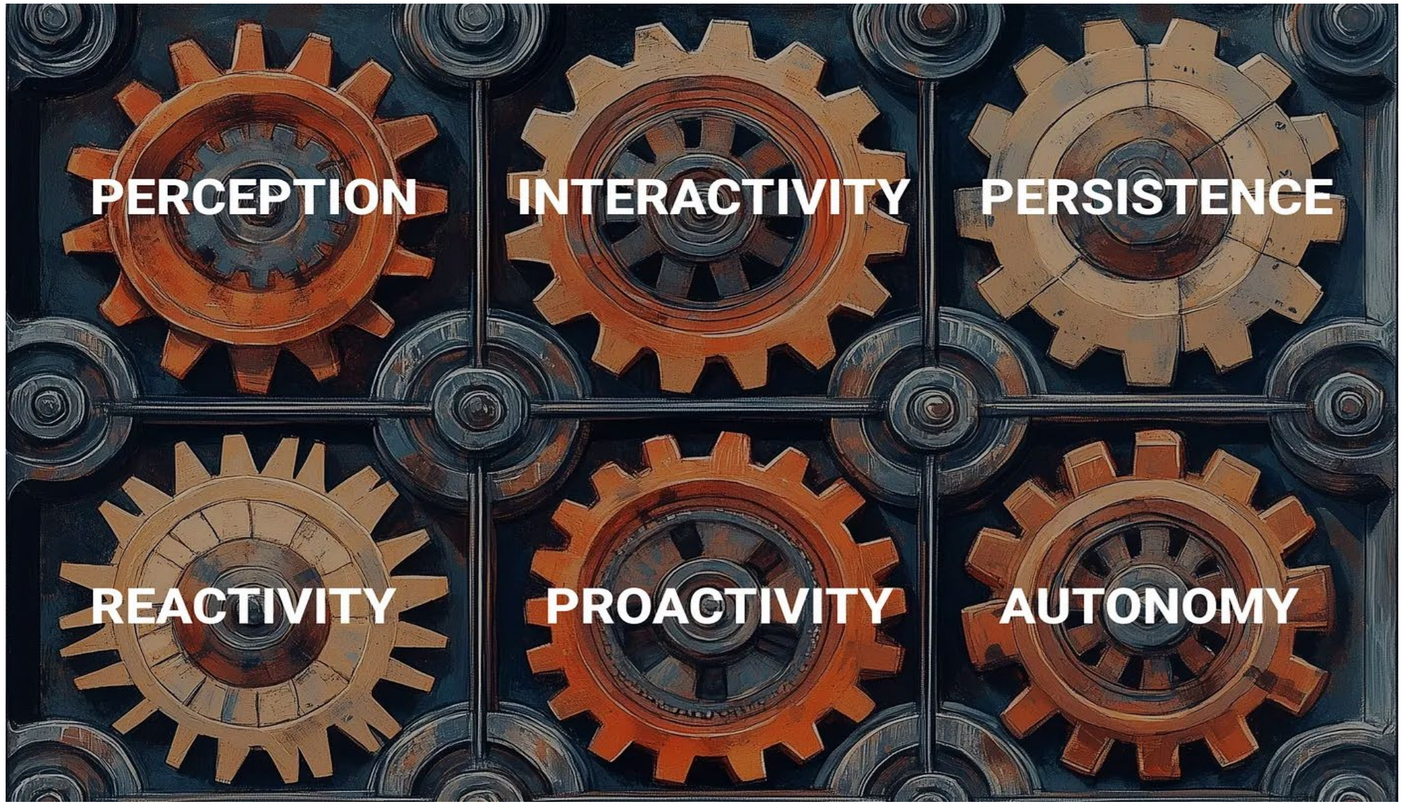
2. Defining Future Vision: Agentic AI

Defining Future Vision: Agentic AI

Agentic AI Systems :

- Capable of setting and pursuing complex goals **independently**
- Adaptable with broader **autonomy** and **decision-making** abilities

Capabilities needed for Agentic AI



https://towardsdatascience.com/what-makes-a-true-ai-agent-rethinking-the-pursuit-of-autonomy-547ab54f4995

Capabilities needed for Agentic AI

1. Perception

The ability to sense and interpret environment or data streams

1. Processes multiple data streams simultaneously
2. Identifies patterns in historical data
3. Interprets complex contextual information
4. Basic: Simple text input understanding
5. Advanced: Integration of multiple data sources (travel history, real-time data, weather, events, news)

Capabilities needed for Agentic AI

2. Interactivity

The ability to engage effectively with environment and users

1. Maintains natural conversational flow
2. Asks for clarifications when needed
3. Adapts communication style to context
4. Offers explanations for suggestions
5. Integrates with external systems and services

Capabilities needed for Agentic AI

3. Persistence

The ability to create and maintain long-term memories

1. Builds comprehensive user profiles over time
2. Maintains historical interaction context
3. Updates knowledge base with new insights
4. References past interactions for future decisions
5. Combines read and write capabilities for user data

Capabilities needed for Agentic AI

4. Reactivity

The ability to respond to changes in real-time

1. Monitors multiple data streams continuously
2. Responds to environmental changes promptly
3. Adjusts recommendations based on real-time data
4. Processes and interprets incoming information
5. Makes timely adjustments to existing plans

Capabilities needed for Agentic AI

5. Proactivity

The ability to anticipate needs without explicit prompting

1. Anticipates potential issues or needs
2. Offers unsolicited but relevant suggestions
3. Flags important deadlines or requirements
4. Makes context-aware recommendations
5. Balances initiative with user autonomy

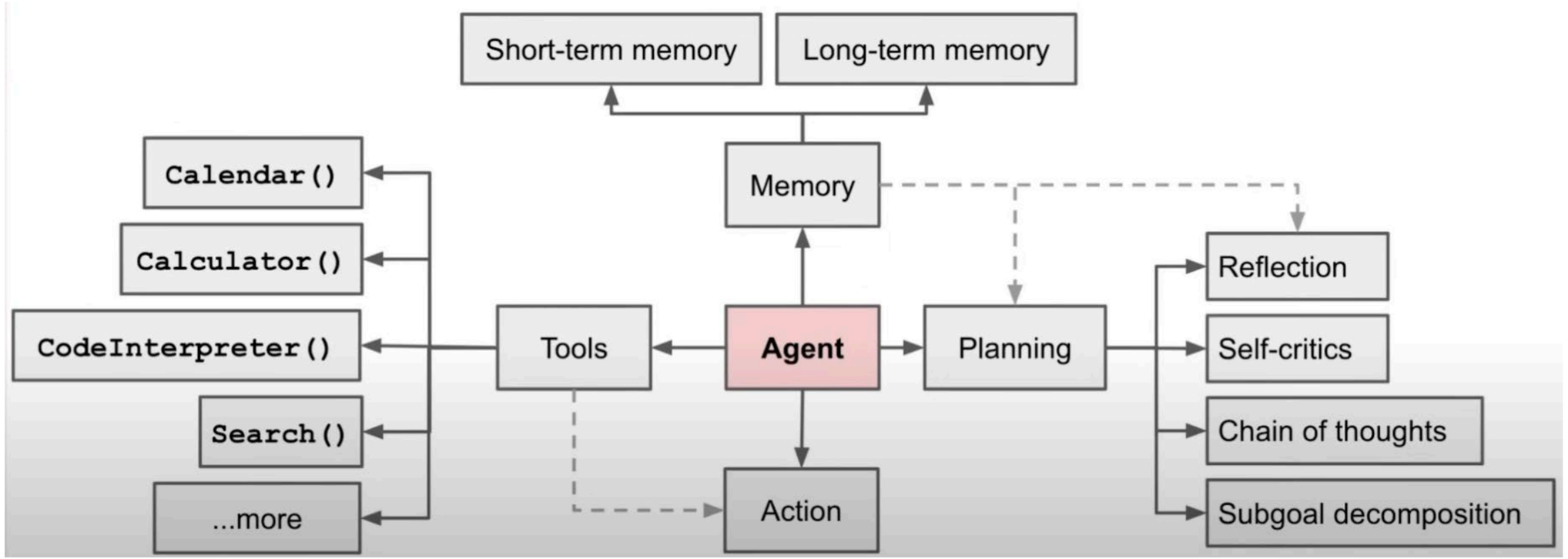
Capabilities needed for Agentic AI

6. Autonomy

The ability to operate independently with defined parameters

1. Controls and allocates resources independently
2. Makes decisions with system-wide impact
3. Operates within defined boundaries
4. Manages complex trade-offs
5. Scales from basic to high-stakes decisions

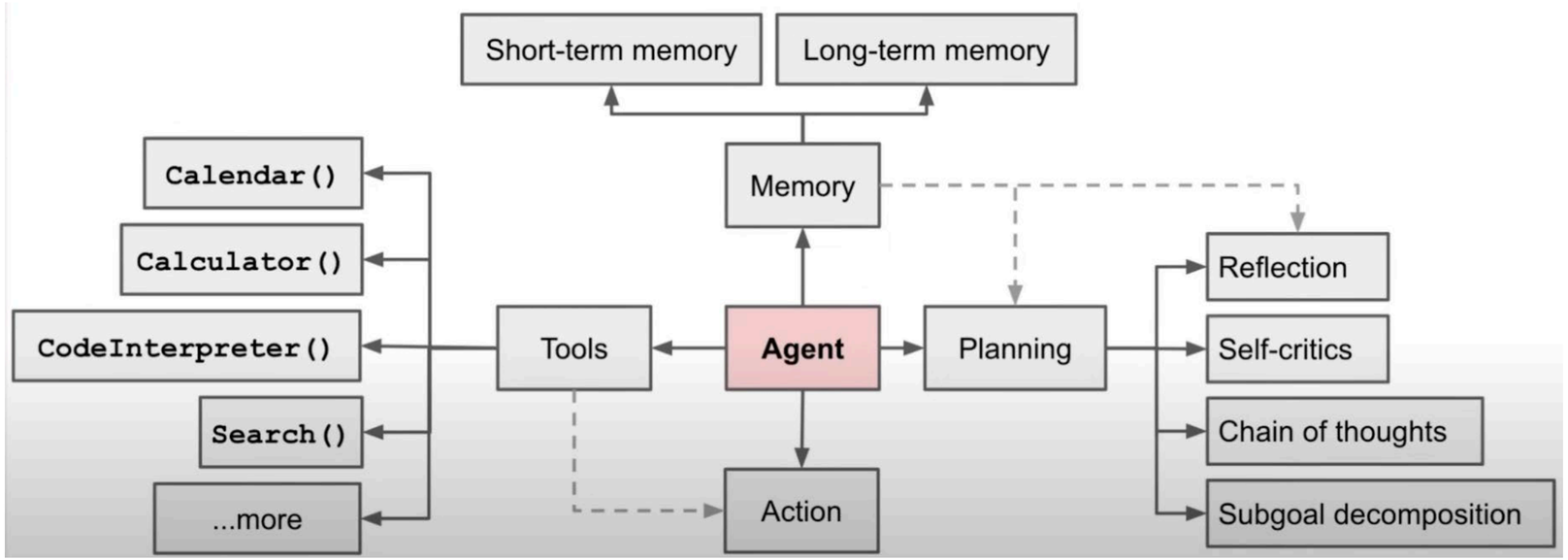
Logical Architecture



Harrison Chase, Co-founder, Langchain

3. WHAT IS POSSIBLE CURRENTLY

What is Possible Currently



- Planning (Decompose Intentions into actions & sequences)
- Memory (Remembering domain specific interactions)

What is Possible Currently

Planning

Language Models don’t do this reliably today

This impacts how we implement Orchestration

Connected with Memory

Memory

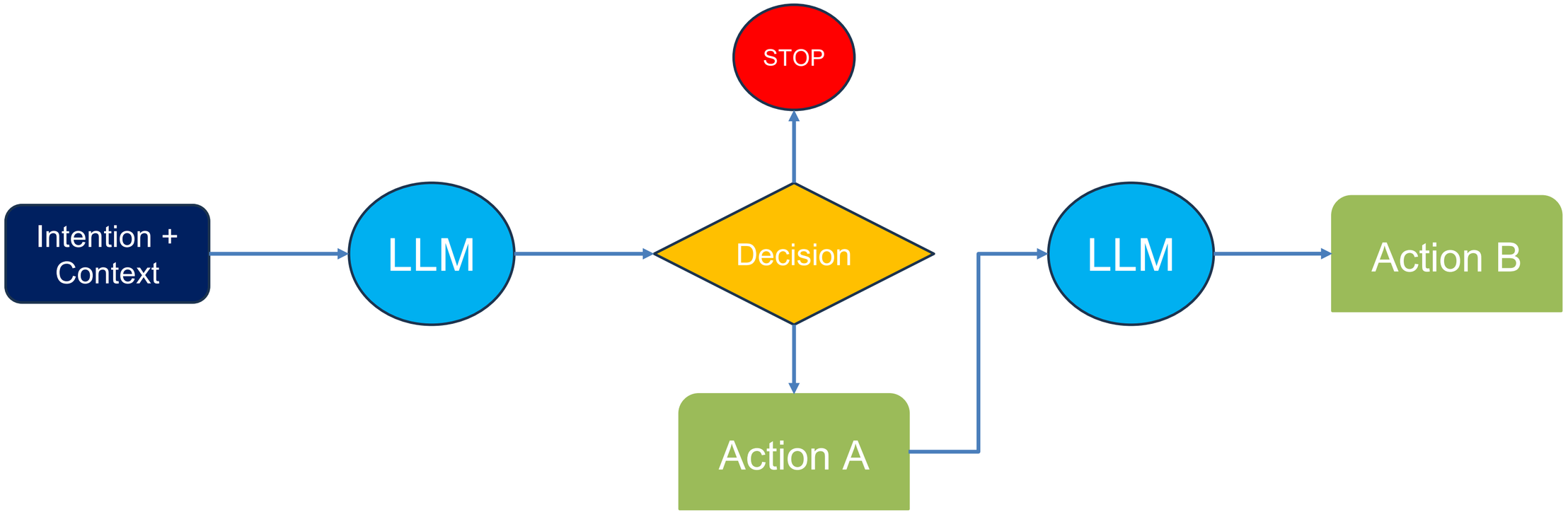
Remembering Facts

Episodic Memory - Events

Ability to update memory during or after the interactions / observations

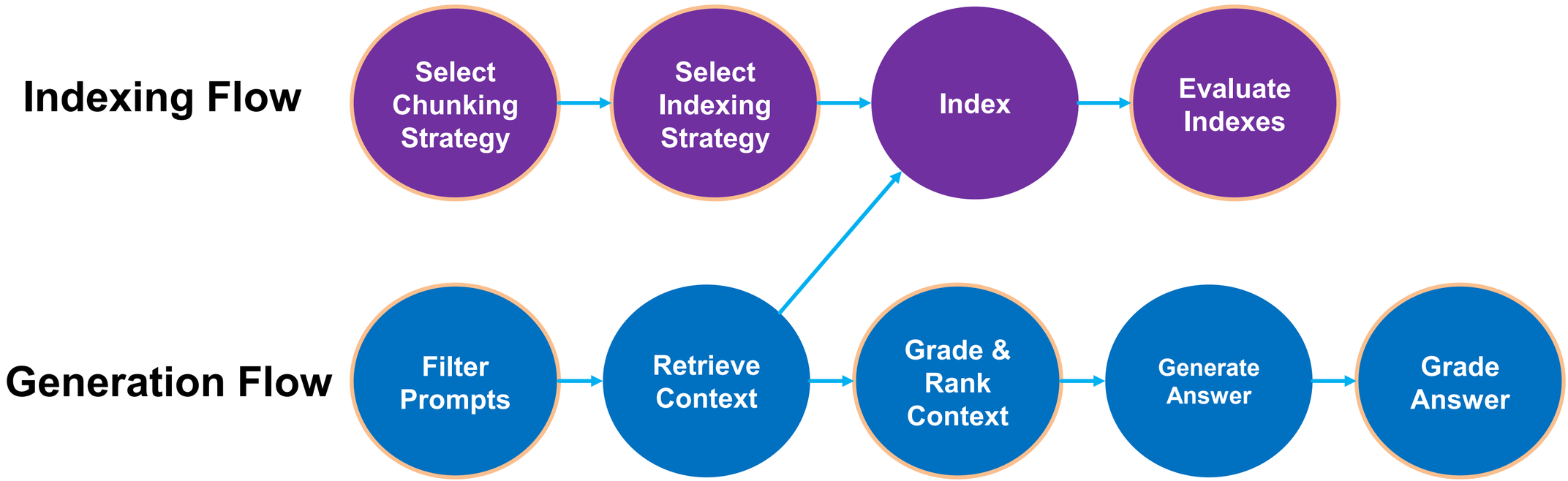
What is Possible Currently

Today, we take over the orchestration with regular consultations with LLMs (by getting the next-immediate action)



4. ADOPTING AGENTIC MODEL IN RAG

Adopting Agentic Model in RAG

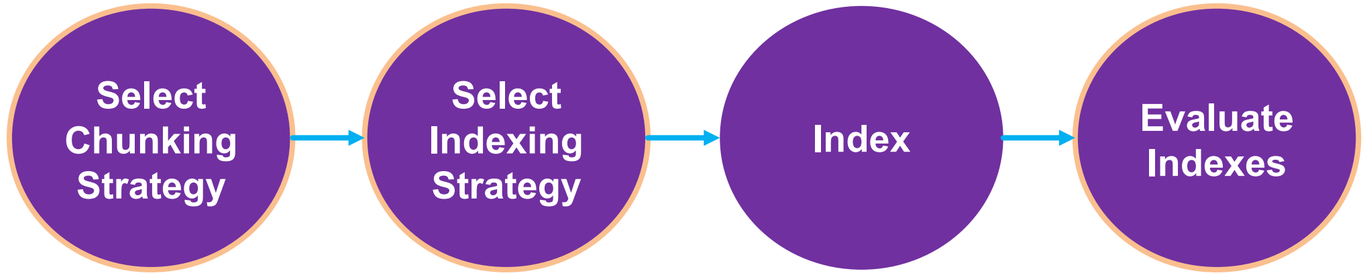


Adopting Agentic Model in RAG

- Context-aware
- Hierarchical
- Semantic
- Sliding Window
- Fixed Length

- Context Precision
- Context Recall

Indexing Flow



- Knowledge Graph
- Hierarchical
- Dense Vector
- Inverted Index

Adopting Agentic Model in RAG

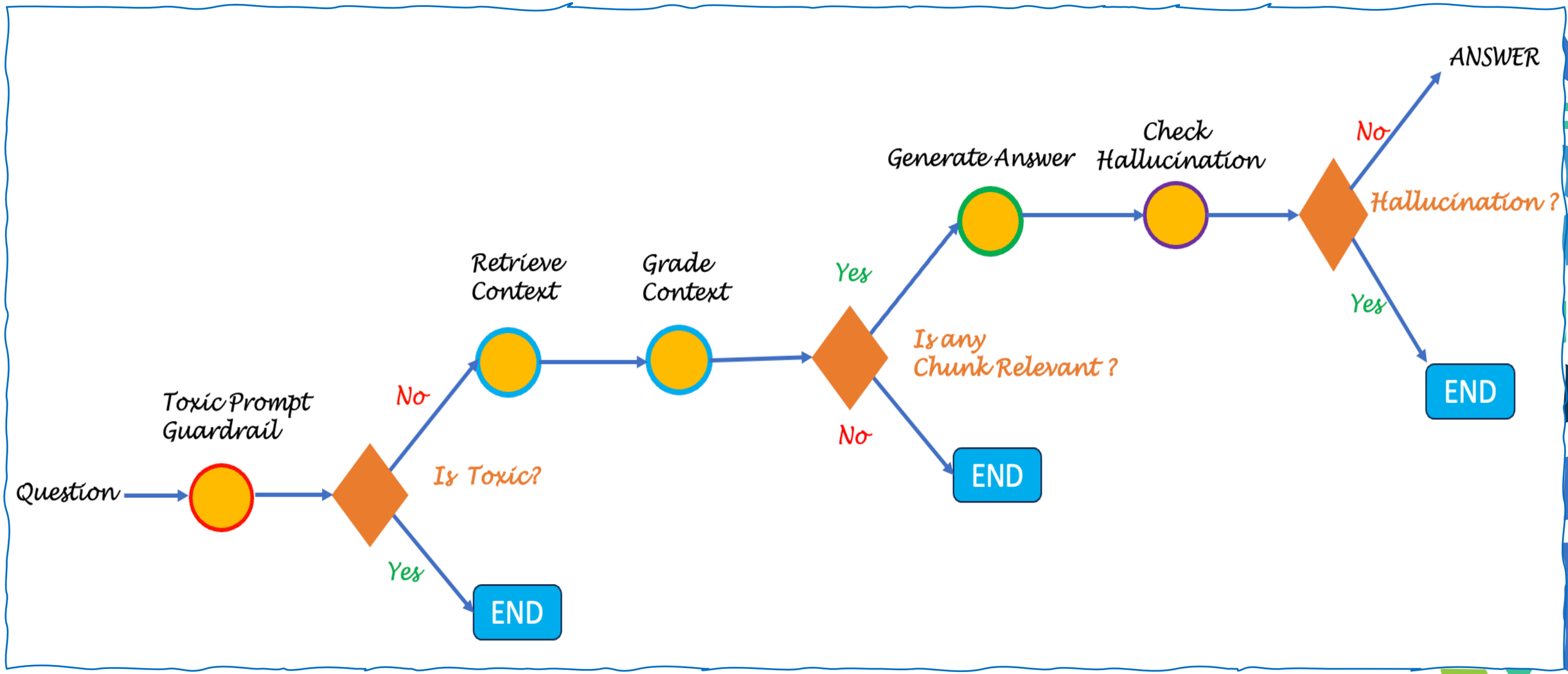
- Purpose-built Models
- LLM-as-a-Judge

- LLM-as-a-Judge
- Faithfulness
- Answer Relevancy

Generation Flow

- LLM-as-a-Judge

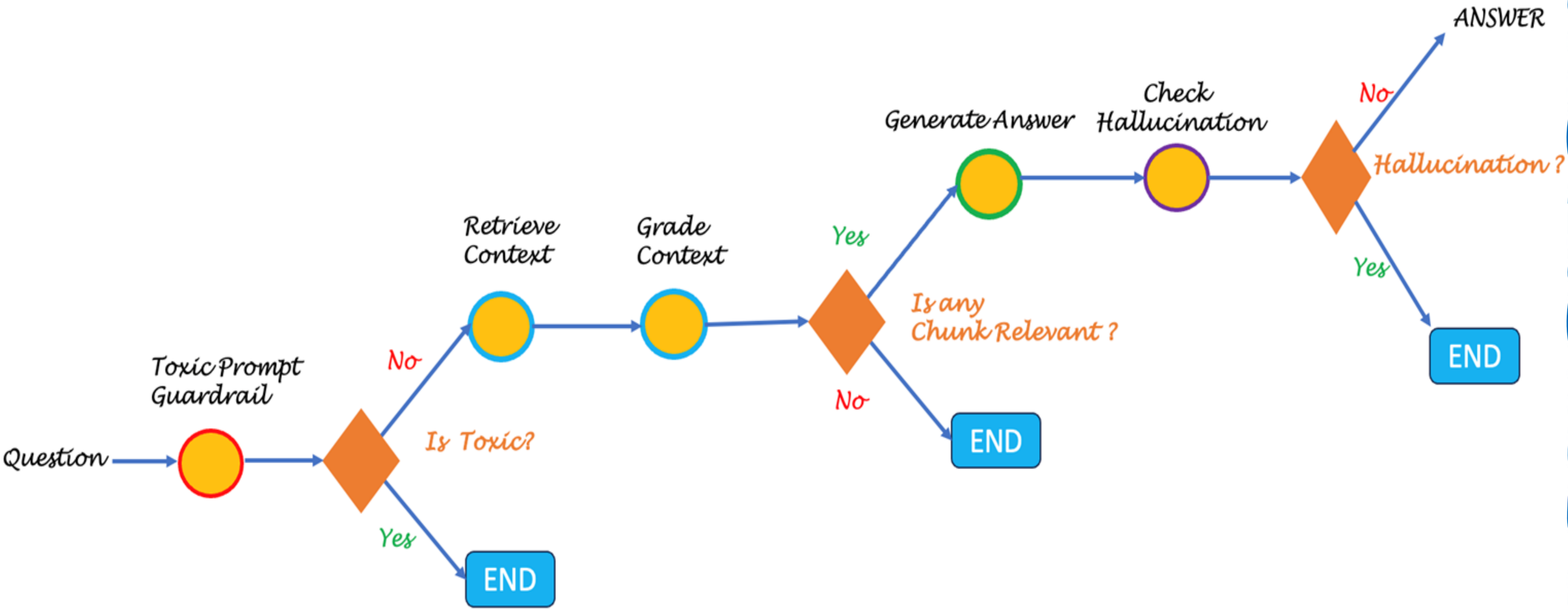
Agentic RAG Implementation Flow



5. DEMO

Summary

Agentic RAG Implementation Flow



Summary

Agentic AI Capabilities

- Perception
- Interactivity
- Persistence
- Reactivity

References

- https://github.com/bhuvana-s/agentic_rag
- Building Agentic RAG Systems with Langgraph
- What Makes a True AI Agent? Rethinking the Pursuit of Autonomy

THANK YOU!

Bhuvaneswari Subramani

@installjournal

https://www.linkedin.com/in/bhuvanas/

https://bhuvana.pro