

RAG

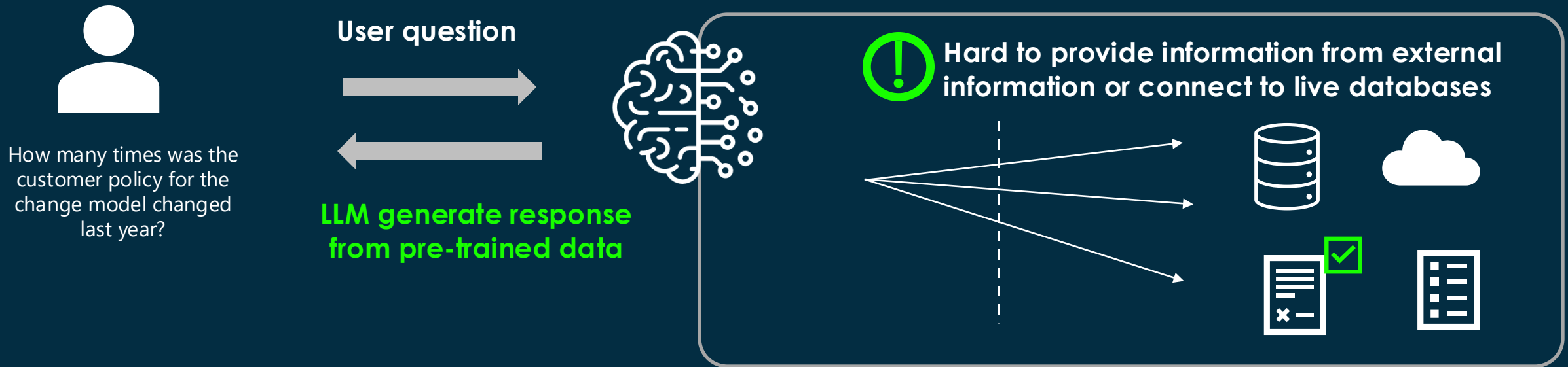
How to accelerate your AI Agent power boost

04/17/2025

lhnaee Choi, Advisory Solution Consultant, Intelligent Automation – AI Specialist

LLMs **generate** responses based on patterns learned from pre-trained data

In general, LLMs don't actively traverse data sources in real-time



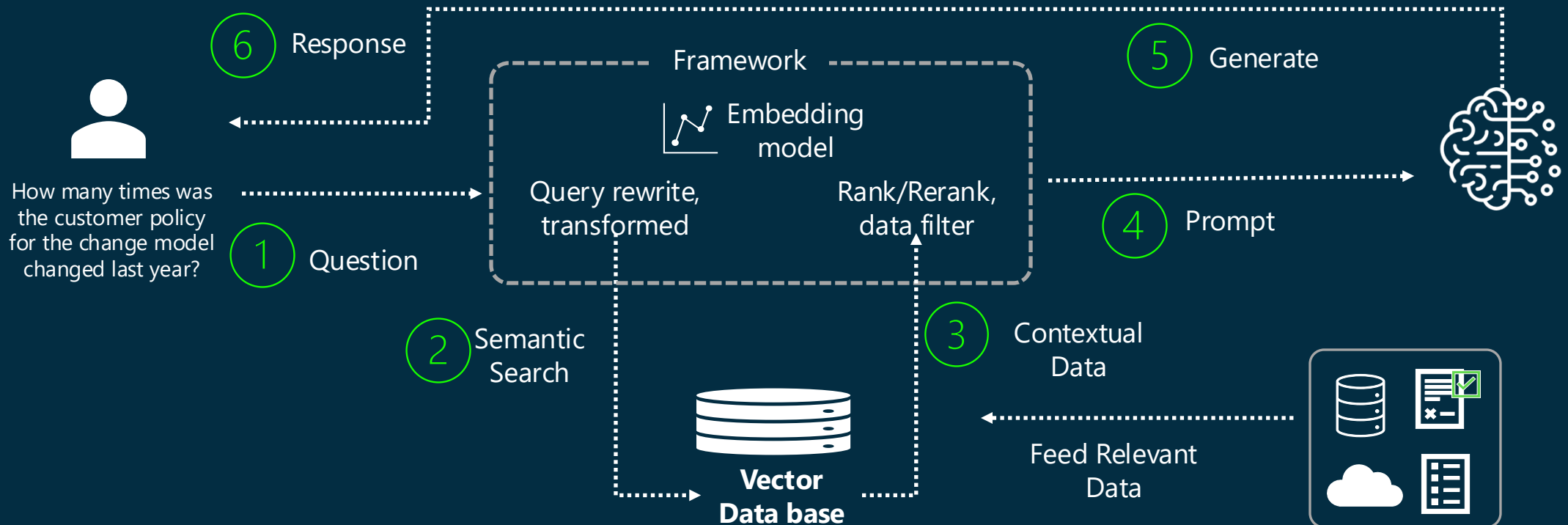
- when the **query** may be interpreted differently from user's intention
- when **no pretrained** data patterns exists



Hallucination

RAG enables LLMs to retrieve and incorporate data in real-time

User question transformed to **retrieve** relevant data → LLMs get contextualized with **augmentation** with relevant data → **Generate** the response accurately and in a timely manner



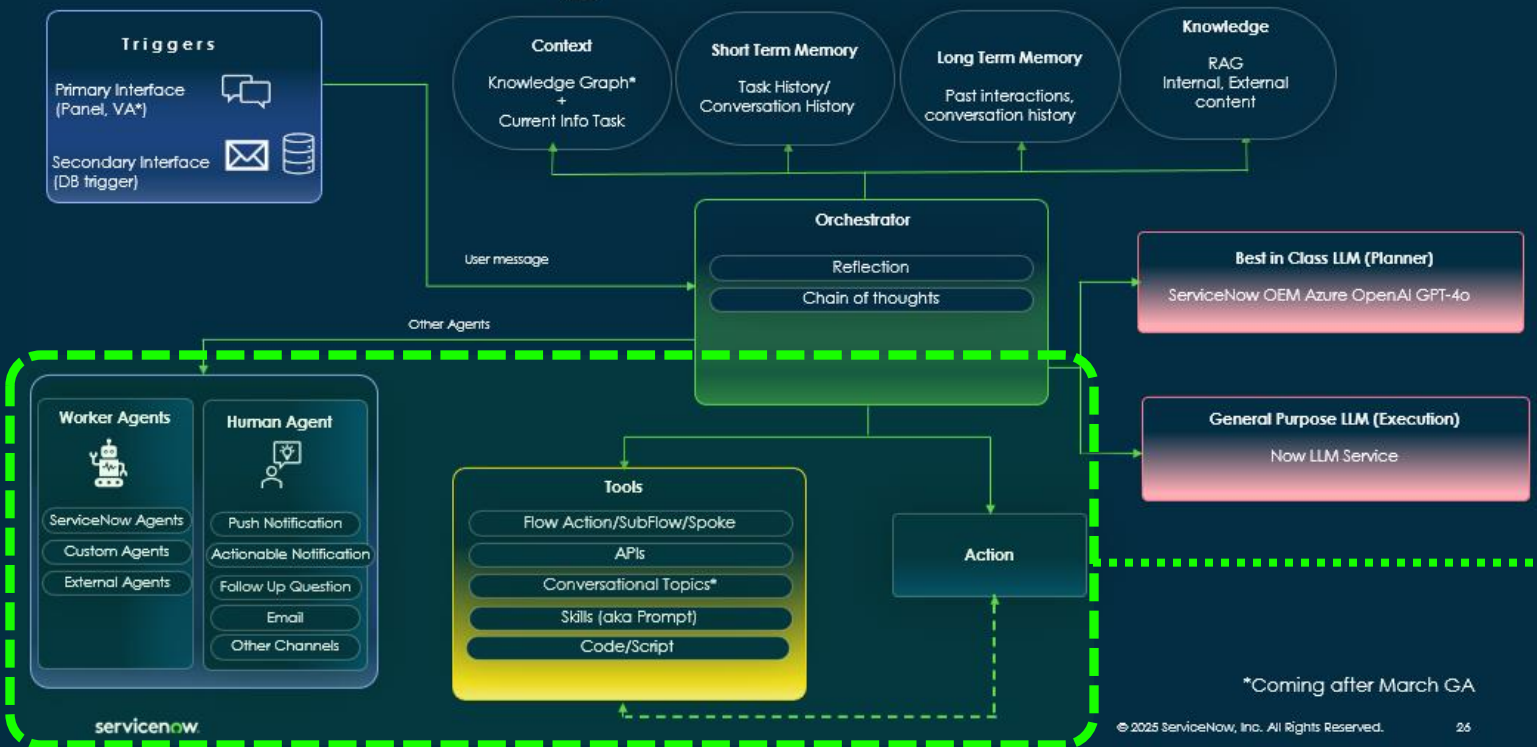
RAG is one example of a **Compound AI System**

RAG

Type of system	Components
Prompt engineering	<ul style="list-style-type: none">• Prompt generation logic• LLM
Unstructured data RAG	<ul style="list-style-type: none">• LLM• Retrieval system
Structured data RAG	<ul style="list-style-type: none">• LLM• Data API e.g., Databricks Online Table• Text-to-SQL engine
Agent-based Chain	<ul style="list-style-type: none">• Function-calling capable LLM• RAG chain• Orchestration chain
Orchestration Chain	<ul style="list-style-type: none">• Function-calling capable LLM• API services

We are using **RAG** as a tool for AI Agent

AI Agents architecture



- LLM's internal knowledge might be outdated or insufficient.
- **RAG** is essential for leveraging internal, external knowledge sources to enhance accuracy and relevancy in generating AI agent strategies.
- We are using RAG to transform query, retrieve the knowledge and prompting the LLM with relevant context to generate accurate and informed responses.

Now Assist AI Agents Studio provides **RAG** framework

servicenow All Favorites History Workspaces Now Assist AI Agents Studio

Overview Testing Settings

Email Cleaner

Describe and instruct ☒

Add tools and information ☒

Define availability ☐

Add tools and information

Add a single tool or information source for the AI agent to begin working. Additional tools can be added.

Scripts AI Instruction

Use scriptable APIs and backend integration to support the AI agent with scripts.

Name	Execution mode	Display output	Description	Date added	Remove
Distribution list unsubscribe	Supervised	true	This tool unsubscribes users from email distribution lists. This tool is used to resolve the given task number	2025-01-30	

Subflows AI Instruction

Subflows are reusable sequences of processing steps that can be called from within a flow.

Name	Execution mode	Display output	Description	Date added	Remove
Add work note	Supervised	true	This flow adds a work note to the Task being resolved. The work note describes how the steps taken to resolve the Task.	2025-01-30	

Add RAG

Retrieval and incorporation of relevant information from an external source.

Name *

Description * AI instruction

Execution mode * AI instruction

☒ Supervised ☐ Autonomous

Display output * AI instruction

☒ Yes ☐ No

Search profile *

Search sources

Fields returned