

RAG

How to accelerate your AI Agent power boost

04/17/2025

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

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



How many times was the customer policy for the change model changed last year?

User question

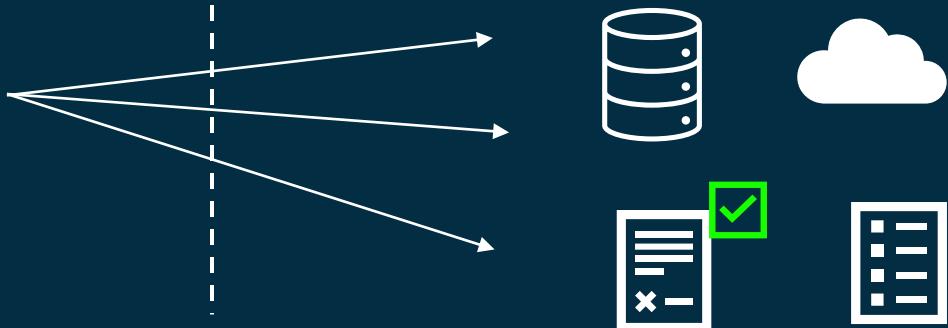


LLM generate response from pre-trained data

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



Hard to provide information from external information or connect to live databases



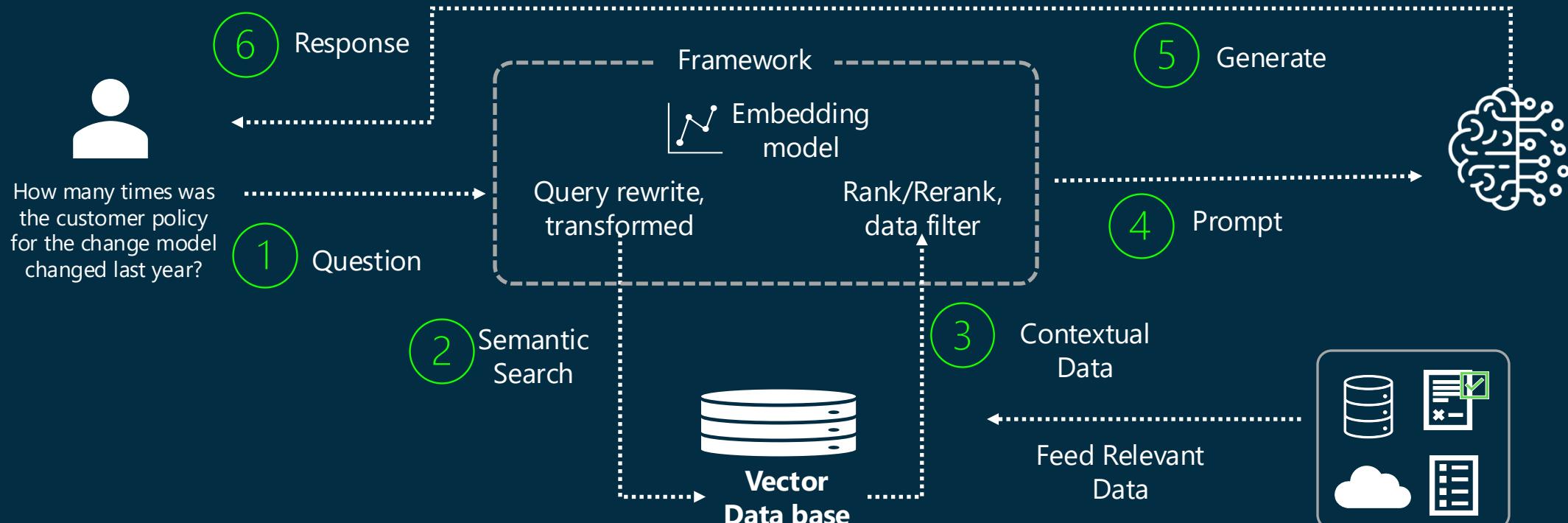
- when the **query** may be interpreted differently from user's intention
- when **no pretrained data patterns** exists



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

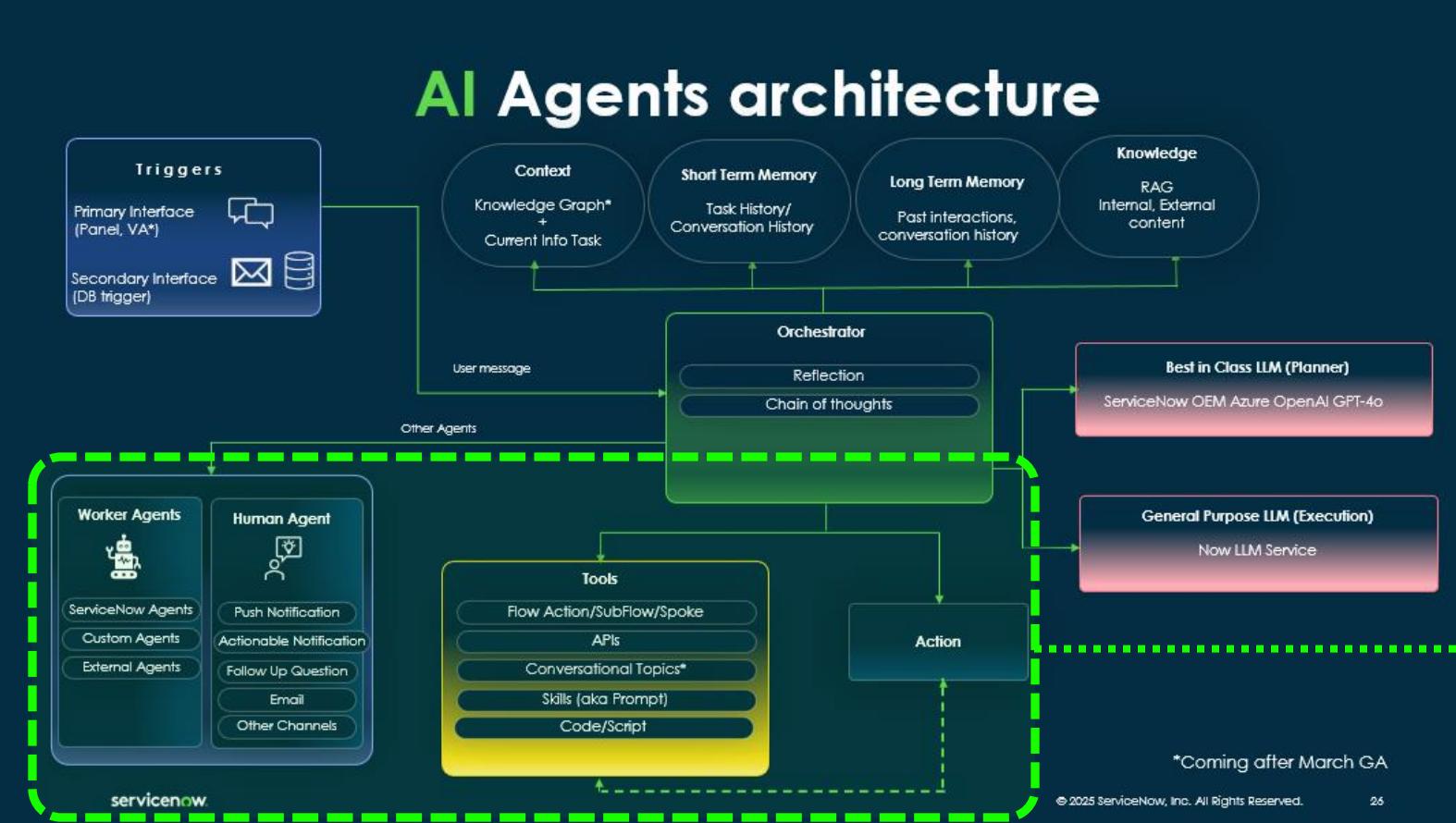


RAG is one example of a Compound AI System

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



- LLM's internal knowledge might be outdated or insufficient.
- **RAG** is essential for leveraging internal, external knowledge sources to enhance 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

The screenshot shows the ServiceNow Now Assist AI Agents Studio interface. On the left, there's a sidebar with options: 'Describe and instruct' (selected), 'Add tools and information' (highlighted with a green dashed box), and 'Define availability'. The main area is titled 'Email Cleaner' and contains sections for 'Add tools and information' and 'Subflows'.

Add tools and information: This section has a sub-section for 'Scripts' (AI Instruction). It lists a single entry: 'Distribution list unsubscribe' (Name), 'Supervised' (Execution mode), 'true' (Display output), and a detailed description. A modal window is open over this section, titled 'Add tool', with a dropdown menu. The 'Retrieval-Augmented Generation (RAG)' option is highlighted with a blue selection bar and is being pointed to by a green arrow from the top right.

Subflows: This section lists a single entry: 'Add work note' (Name), 'Supervised' (Execution mode), 'true' (Display output), and a detailed description. It includes columns for 'Date added' (2025-01-30) and 'Remove'.

Right side panel: A large modal window titled 'Add RAG' is open. It has fields for 'Name' (mandatory), 'Description' (AI instruction), 'Execution mode' (Supervised selected), 'Display output' (Yes selected), 'Search profile' (mandatory), 'Search sources', and 'Fields returned'.