servicenow

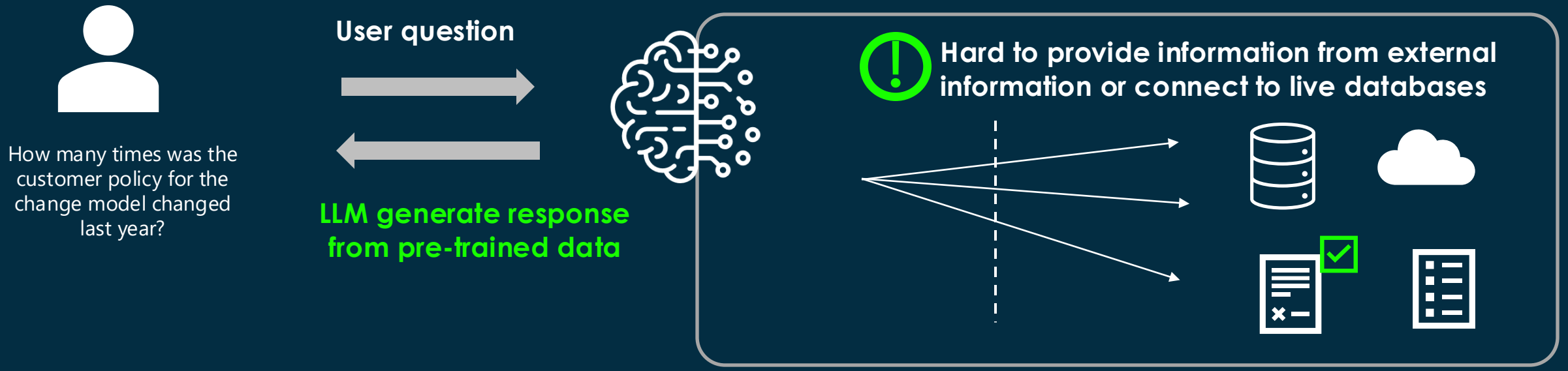
RAG

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

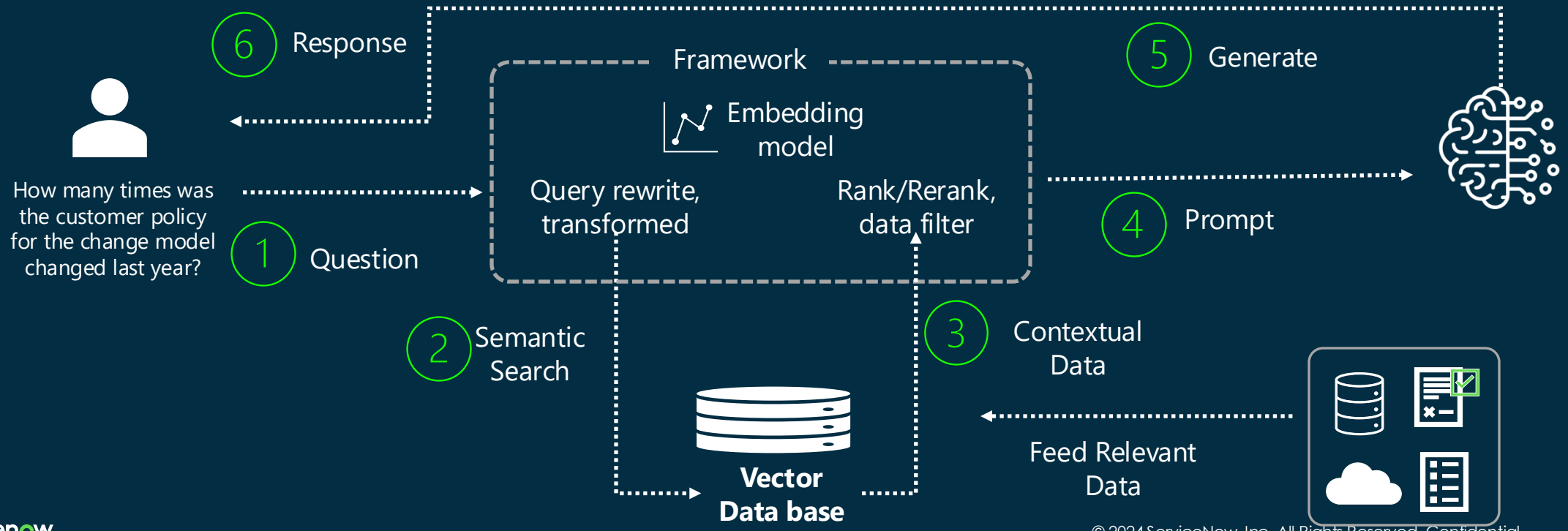


- when the query may be interpreted differently from user's intention
- when no pretrained data patterns exists

→ **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**

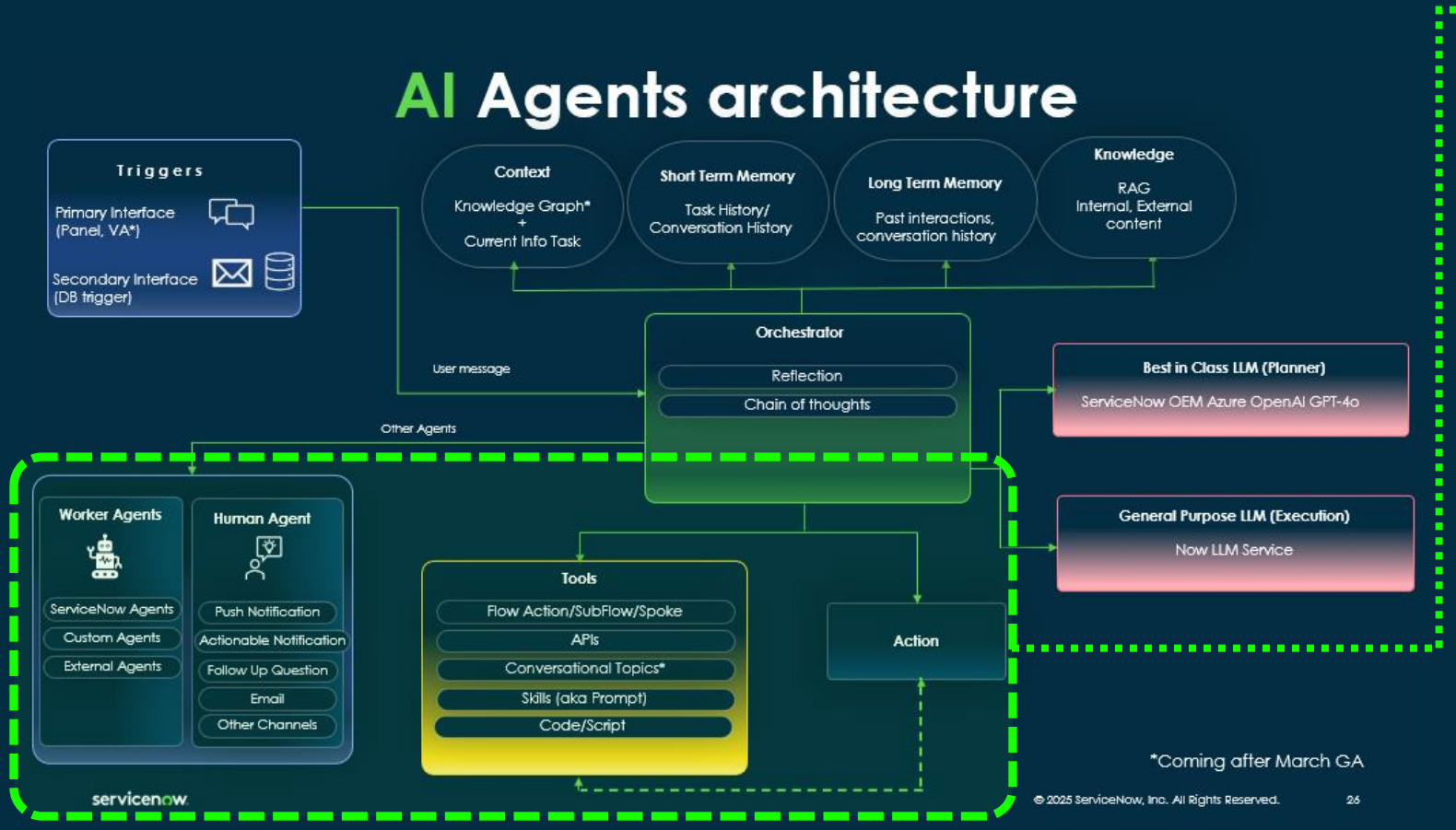


RAG is one example of a Compound AI System

RAG

Type of system	Components
Prompt engineering	• Prompt generation logic • LLM
Unstructured data RAG	• LLM • Retrieval system
Structured data RAG	• LLM • Data API e.g., Databricks Online Table • Text-to-SQL engine
Agent-based Chain	• Function-calling capable LLM • RAG chain • Orchestration chain
Orchestration Chain	• 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- LLM's internal knowledge might be outdated or insufficient.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