BotBuddy 2.0

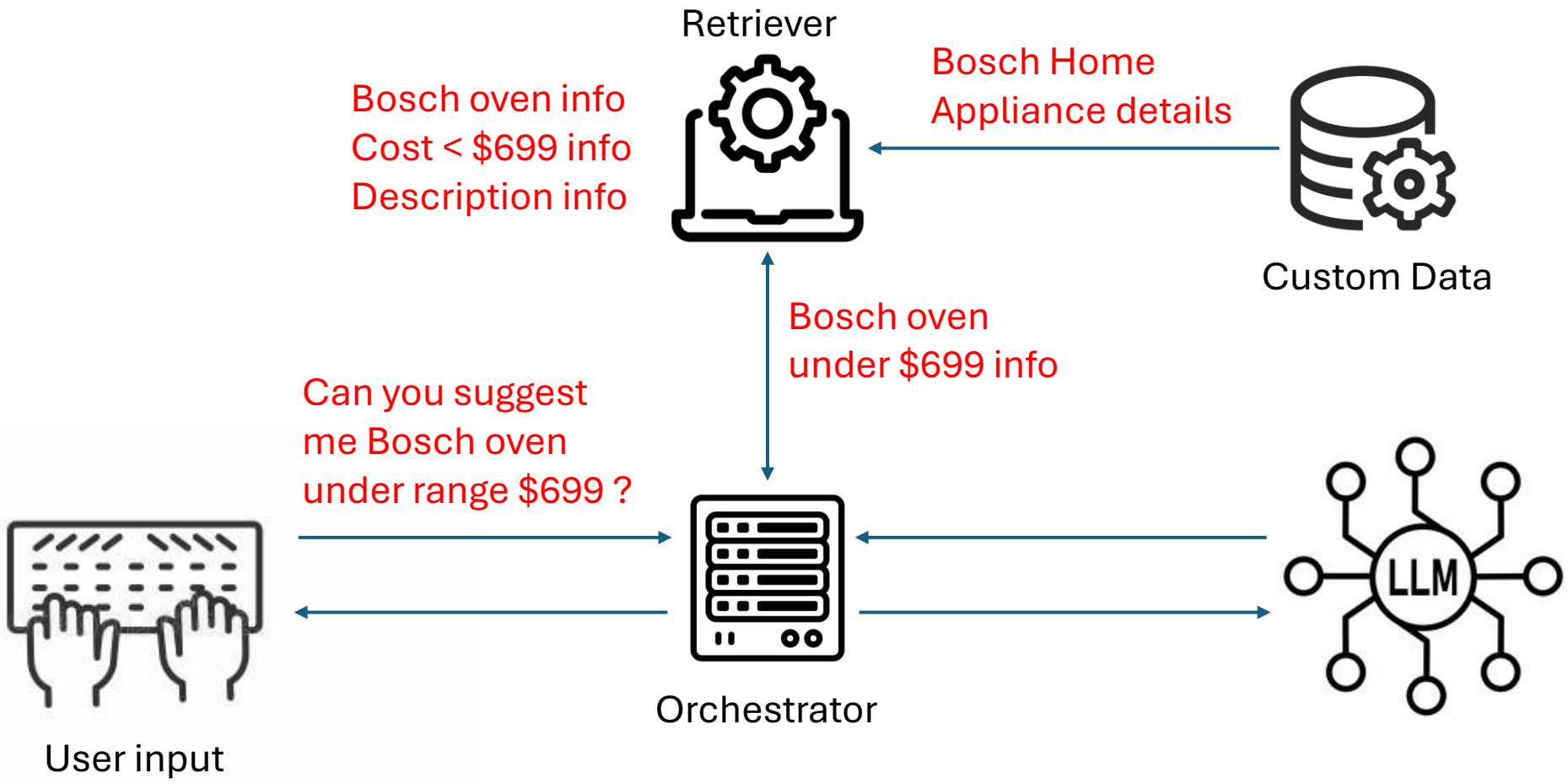
Key Updates

Chat Context Awareness
Remembers the ongoing conversation for smooth, coherent replies across multiple messages.

Reduced Hallucination
Responses are grounded in real data using Retrieval-Augmented Generation (RAG), minimizing fabricated or misleading information.

Retrieval-Augmented Generation (RAG)
Integrates with Azure Cognitive Search to fetch relevant data from a knowledge base for more accurate answers.

Data Privacy
Ensures sensitive data is handled securely — no conversations are stored or shared outside your cloud environment.



Model used:

- gpyt-4o-mini
- Text-embedding-ada-002

BotBuddy 1.0 → BotBuddy 2.0 - - - - - - - - → BotBuddy 3.0

+ve
- Flow + Generative based response
- Deployed on Microsoft Co-Pilot Studio
- Easy to implement (LCNC)

-ve
- Less Generative Content ~ 10%
- Often Hallucinates
- High response time
- No evaluation
- Low customization

+ve
- Generative content >50%
- Deployed on Microsoft AI Foundry
- LLM are deployed on secure server which solves data privacy issue.
- Possibility to add bot evaluation
- Low response time
- Highly customizable

-ve
- Can’t handle specific user story.

- Multi Agent model to handle more specific user story.
- Connectivity with MCP server