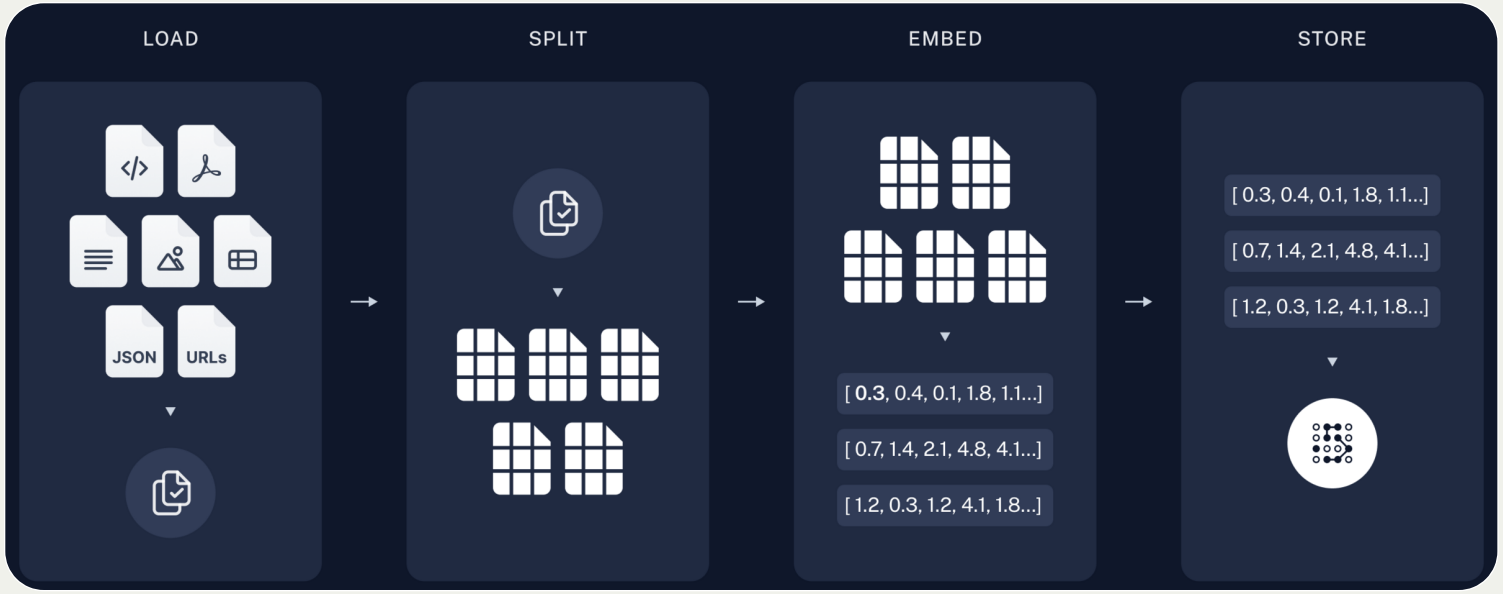
From RAG to Agentic system

Vladimir Ageev

Intro

What is a RAG?

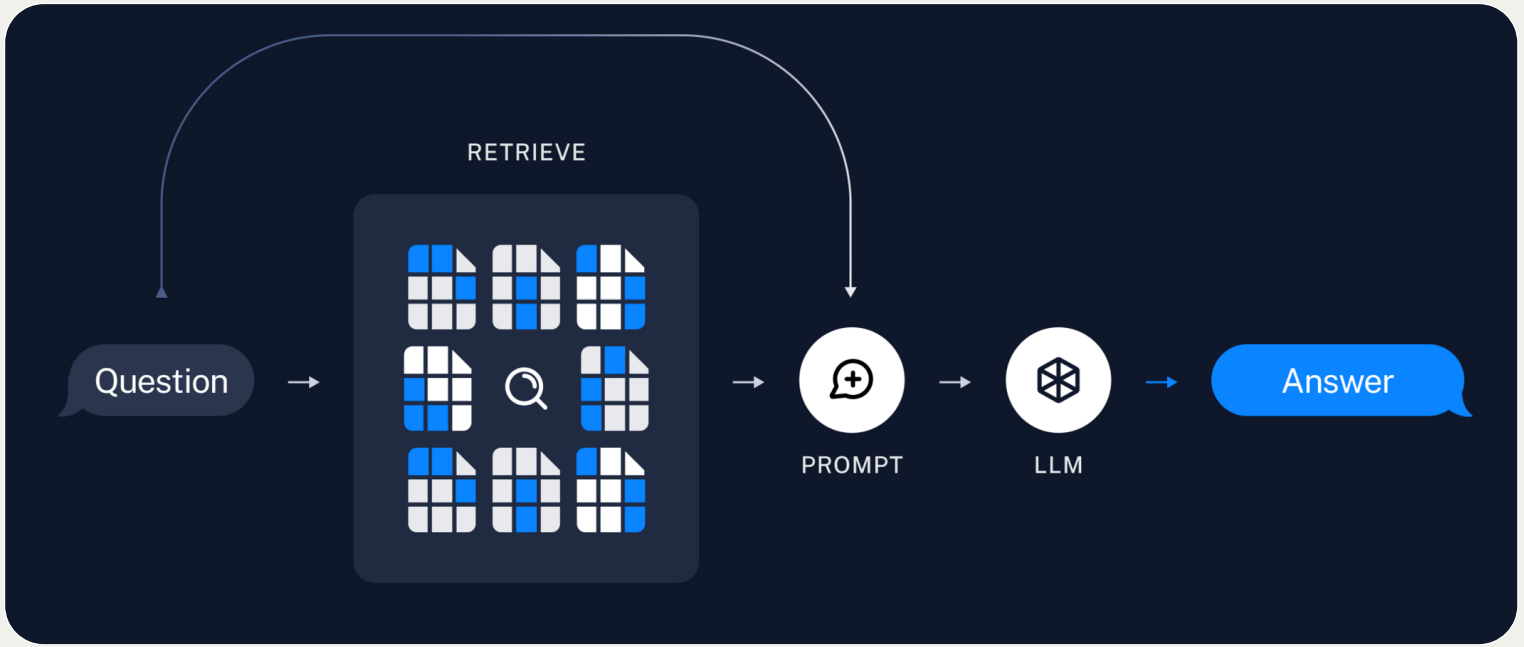
1. **Indexing**
   parsing documents
   chunking
   encoding
   indexing in db



images from: https://python.langchain.com/docs/tutorials/rag/

What is a RAG?

1. Indexing stage
2. **Retrieval & Generation**
   LLM answers question

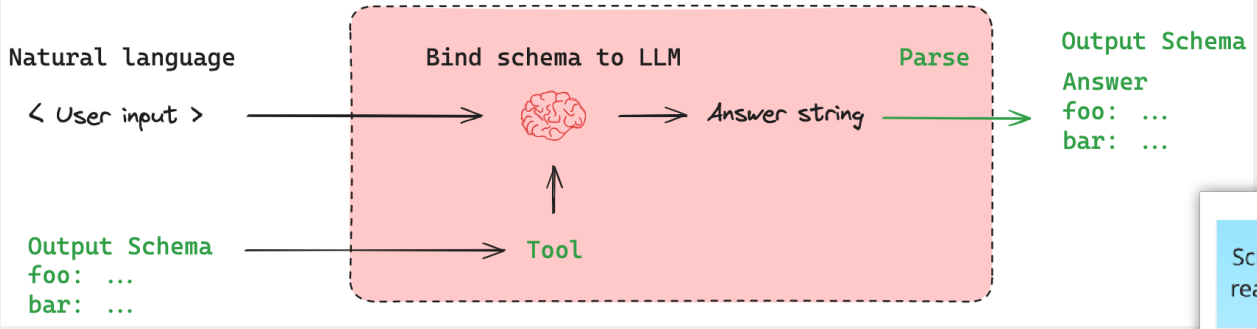


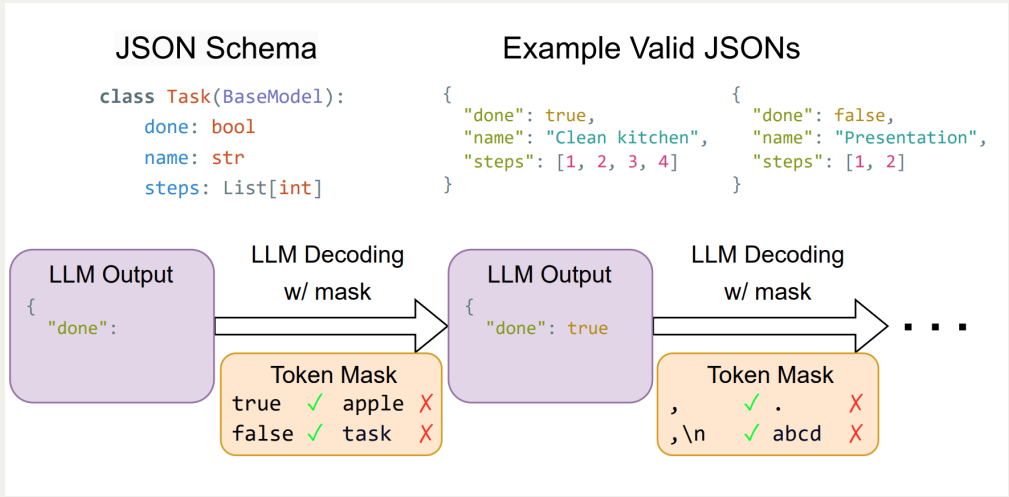
images from: https://python.langchain.com/docs/tutorials/rag/

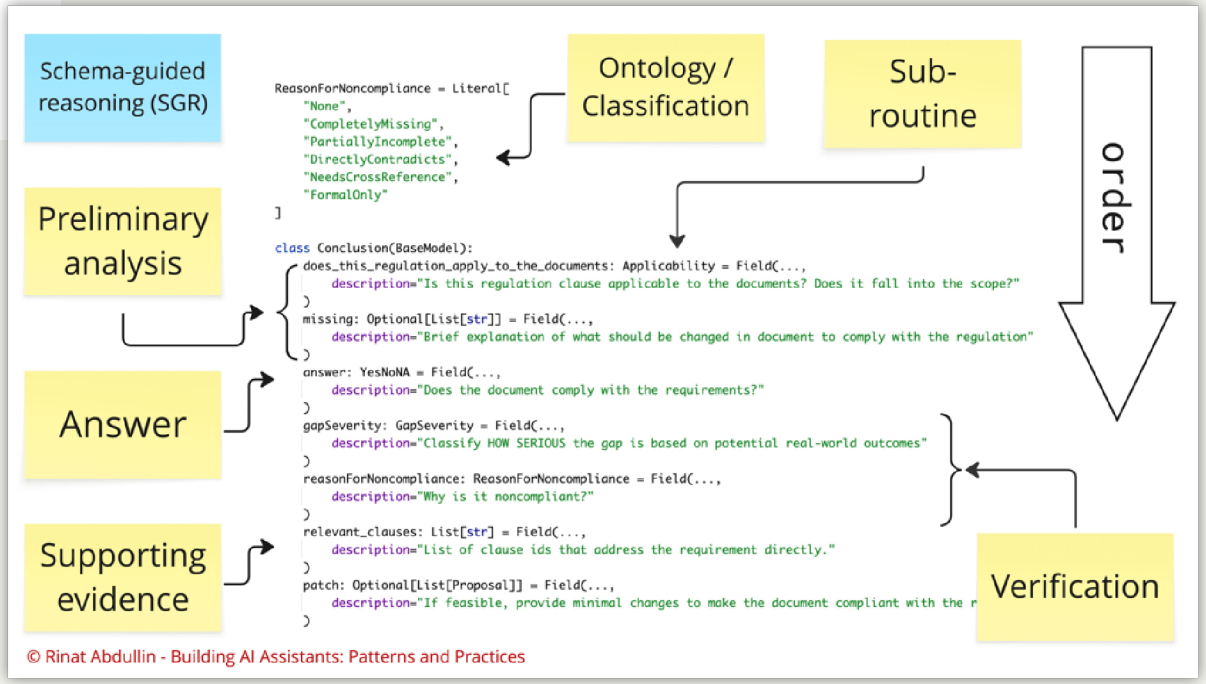
What is missing

- **Up-to-date information beyond context:**
  - attach new sources, browse web
- **Memory and state:**
  - continue conversation or update internal parameters
- **Multi-turn:**
  - run retrieval multiple times with different queries on the same topic

Structured output & Schema Guided Reasoning





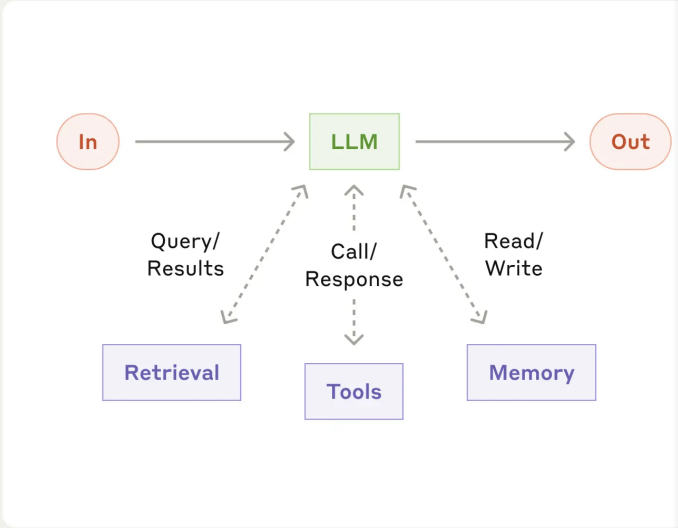


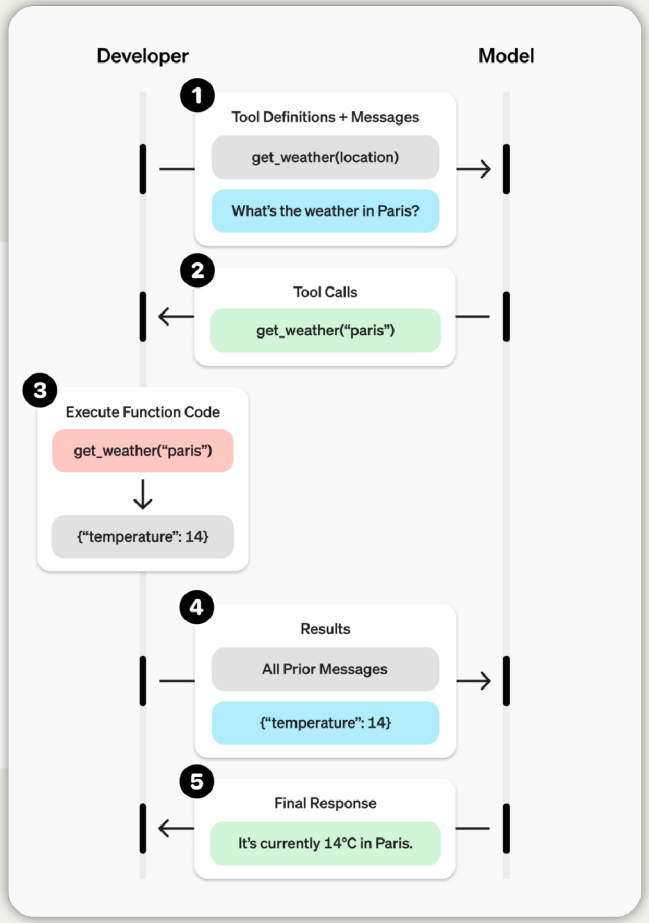
images from: https://python.langchain.com/docs/concepts/structured_outputs/ https://blog.mlc.ai/2024/11/22/achieving-efficient-flexible-portable-structured-generation-with-xgrammar https://abdullin.com/schema-guided-reasoning/

Tools & Function calling

Structured output enables
more stable function calling

Input: tool definitions + schemas
Output: selected tool + arguments



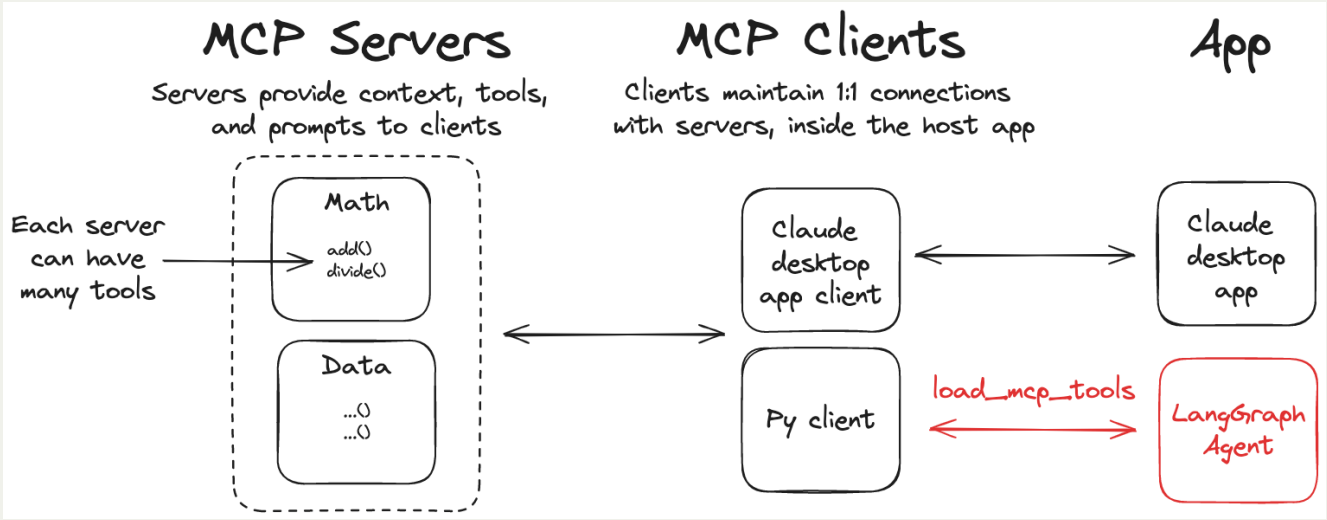


images from: https://platform.openai.com/docs/guides/function-calling#the-tool-calling-flow

Tools: Model Context Protocol

Proposed way for services to expose **tools** to LLMs

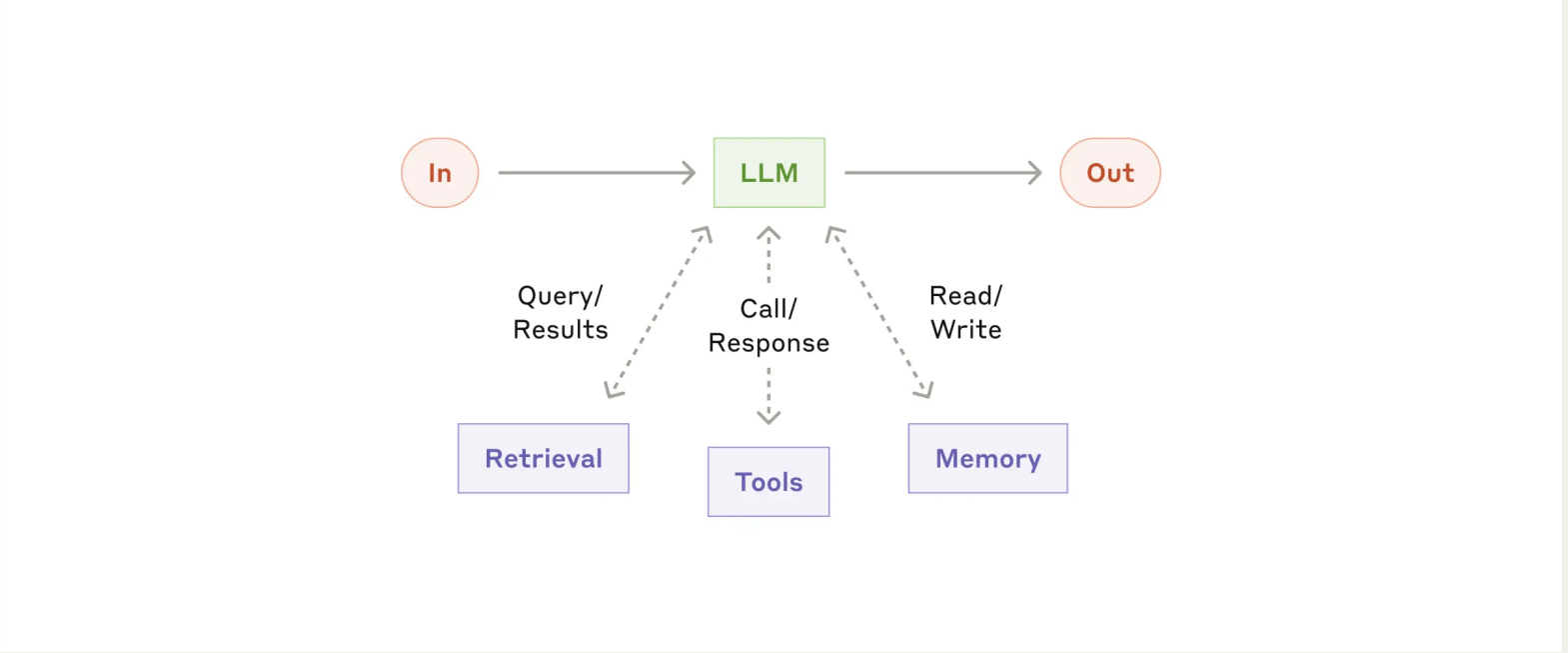
```python
@mcp.tool()
async def get_forecast(latitude: float, longitude: float) -> str:
    """Get weather forecast for a location.

    Args:
        latitude: Latitude of the location
        longitude: Longitude of the location
    """
    # First get the forecast grid endpoint
    points_url = f"{NWS_API_BASE}/points/{latitude},{longitude}"
    points_data = await make_nws_request(points_url)

    if not points_data:
        return "Unable to fetch forecast data for this location."

    # Get the forecast URL from the points response
    forecast_url = points_data["properties"]["forecast"]
    forecast_data = await make_nws_request(forecast_url)
```



images from: https://langchain-ai.github.io/langgraph/concepts/mcp/

What is an Agent?

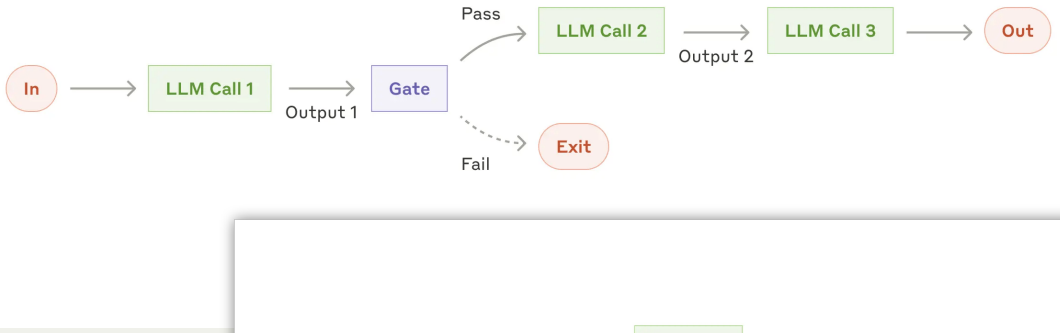
Building block with state and tool usage

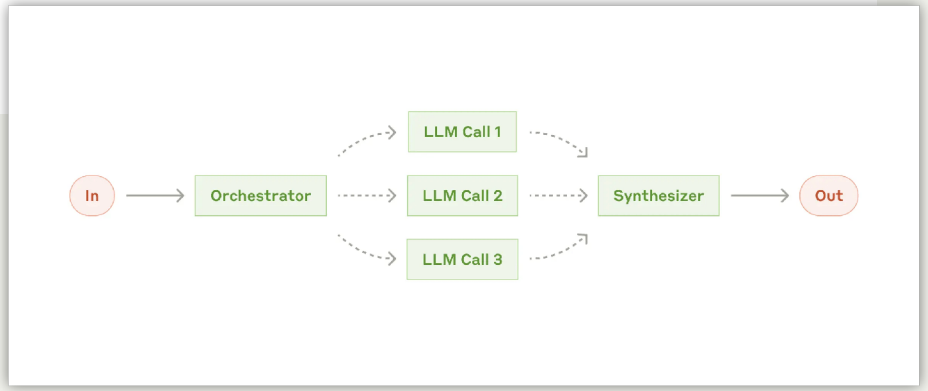


images from: https://www.anthropic.com/engineering/building-effective-agents

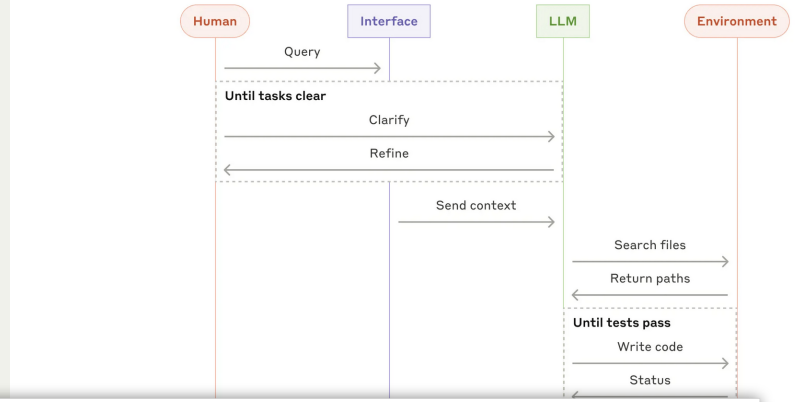
What is an Agent?

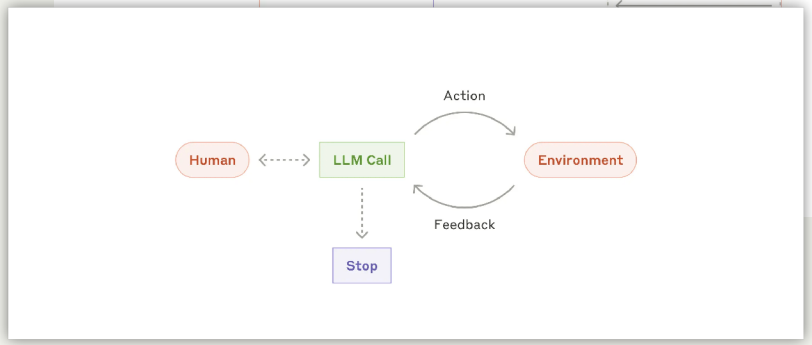
"Workflows"





"Agents"

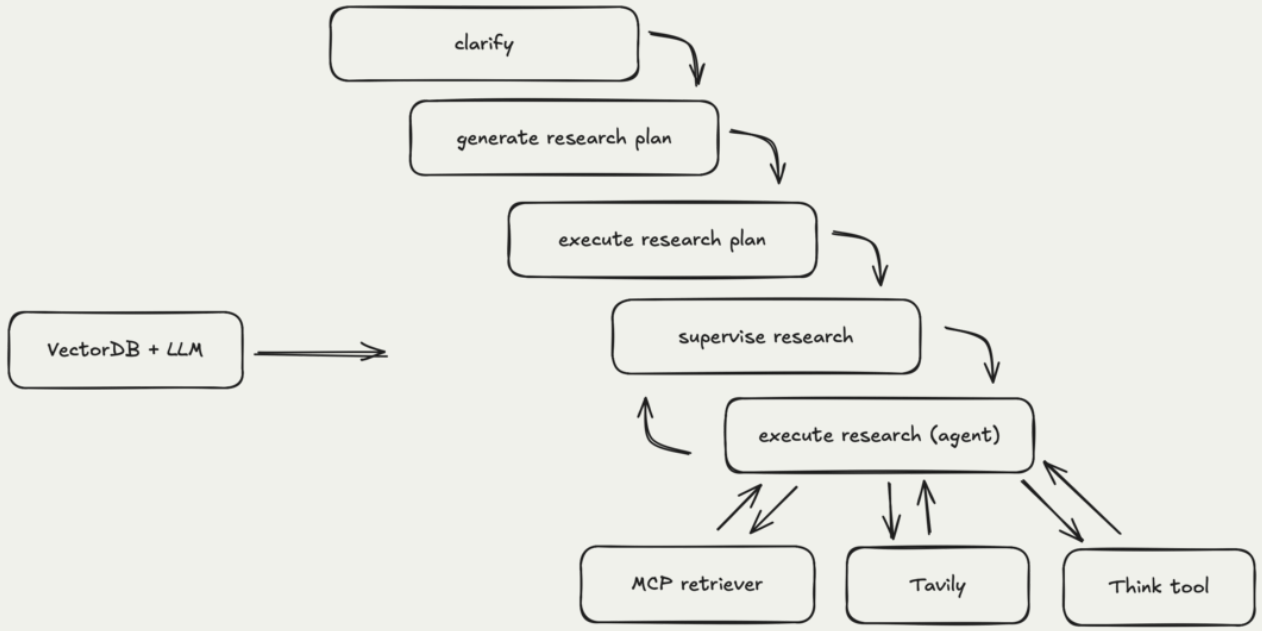




images from: https://www.anthropic.com/engineering/building-effective-agents

We are gonna build

We will walk through:
- structured outputs
- search agent (LLM + Tools)
- retrieval as MCP server
- states and communication between agents



Thank you!

I am

Vladimir Ageev, DS @ EPAM Systems