Last time:

- RAG
  - given a prompt, retrieve relevant docs
  - LLM(prompt; docs) ⇒ response
  - fixed, one-turn control flow

- tool use
  - LLM is made aware of the tools it can use via the prompt
  - at any point during generation, model can issue tool calls
    - special tool tokens
    - tool calls executed as soon as they're produced, output is appended to the generation

↳ agent

↳ generally multi-turn, breaks input task into smaller subtasks that are each executed by tool calls or other LLMs

↳ loops until some goal condition has been satisfied

↳ an LLM controller that makes and revises a plan

↳ multiple LLM workers who execute smaller subtasks

↳ popular agent use cases:

1. coding agent
2. computer-use agent
3. research agents