AI FOR JAVA DEVELOPERS

Spring AI Workshop

Dan Vega ~ Spring Developer Advocate @Broadcom

INTRODUCTION

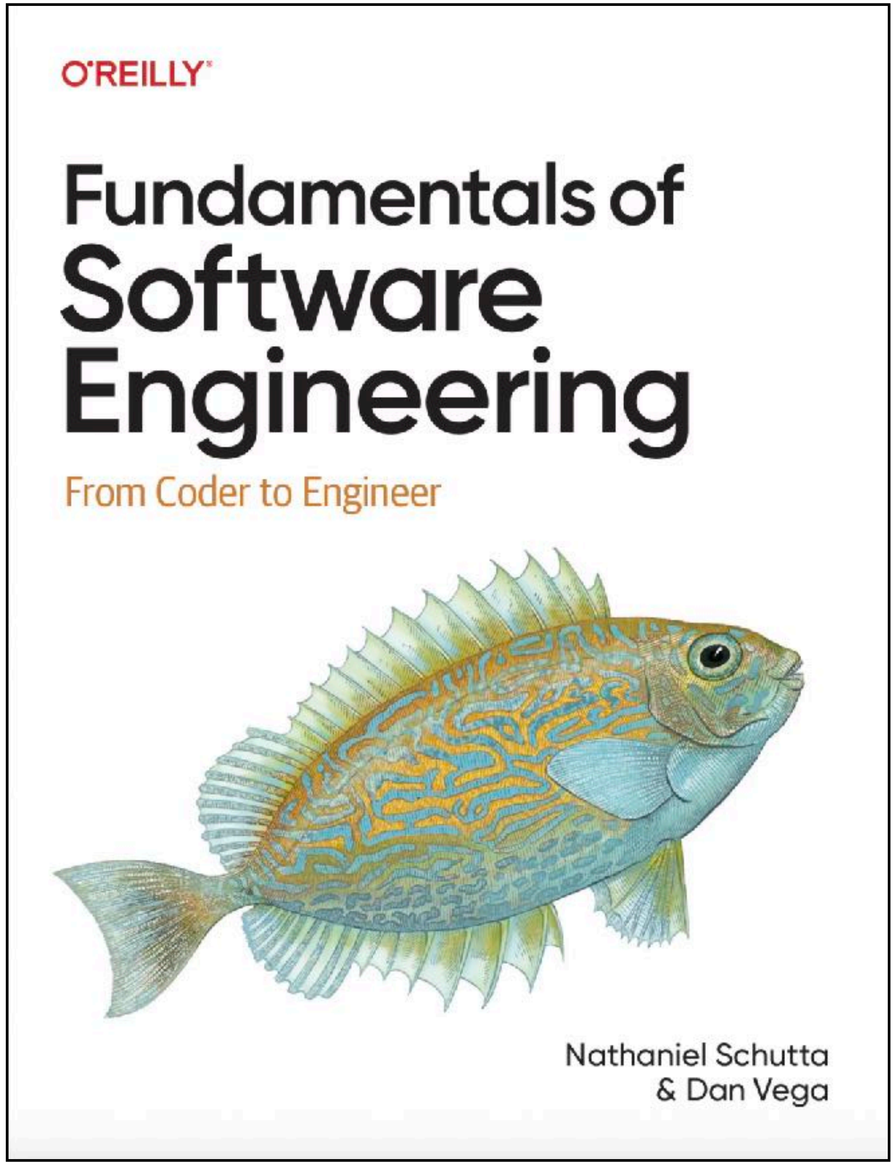
WORKSHOP AGENDA

- 8:00 AM - 10:00 AM :: Introduction & Spring AI
- 10:00 AM - 12:00 PM :: LLM Limitations (Tools, MCP)
- 10 Min Break top of every hour
  - Bio Break, Coffee, Chance to get unstuck

ABOUT ME

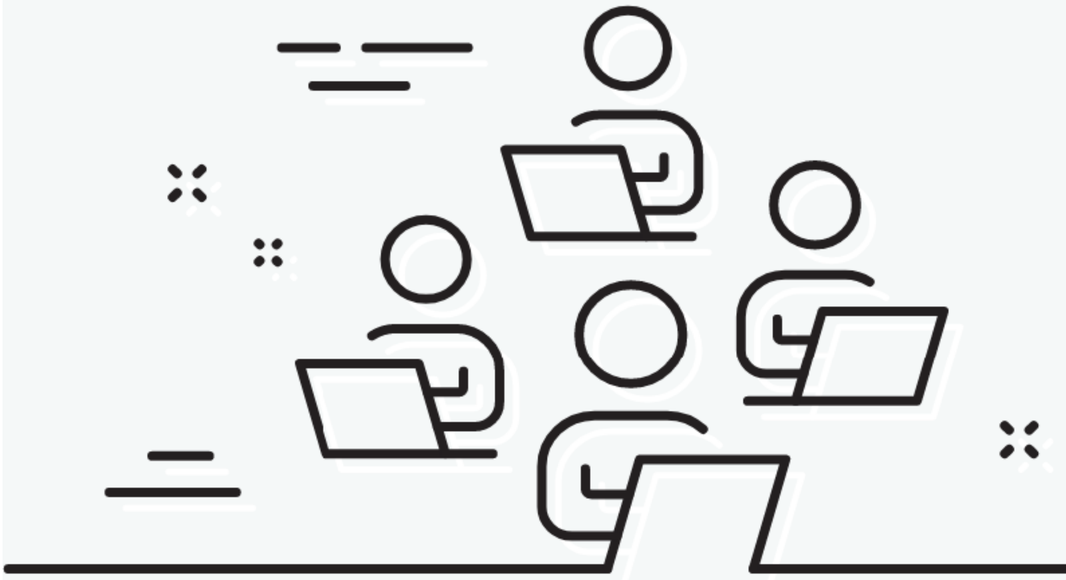
Learn more at danvega.dev

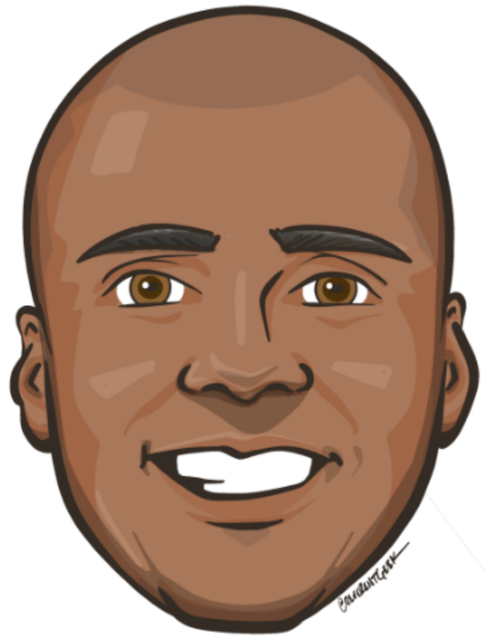
- 👨‍👩‍👧‍👦 Husband & Father
- 🏠 Cleveland
- 🍃 Spring Developer Advocate
- ☕ Java Champion
- 📖 Author
- 👨‍💻 Software Development 23 Years



spring

OFFICE HOURS





https://www.springofficehours.io

PREREQUISITES AND RESOURCES

Prerequisites

- Familiar with the Java programming Language
- JDK 17+ Installed
- Familiar with Spring (but not Spring AI)
- IDE (IntelliJ)

I want to make sure we can all run Java / Spring apps before moving forward

This is a workshop. It is meant to be interactive

If / How are you leveraging AI in your applications today?

https://www.danvega.dev/workshop/ai-for-java-developers

THE JAVA AI OPPORTUNITY

Whenever I learn something new I like to start with

Why are we talking about AI on the JVM?

PYTHON LIBRARIES FOR AI/ML

Core Deep Learning Frameworks

- **TensorFlow** – originally dominant, still heavily used in production/enterprise.
- **PyTorch** – the current favorite for research and a lot of industry adoption.

Classical ML / Data Science Foundations

- **Scikit-learn** – standard for traditional ML (regression, classification, clustering, pipelines).
- **NumPy** – numerical computing foundation (arrays, linear algebra).
- **Pandas** – data manipulation and analysis (DataFrames).

Transformers / NLP

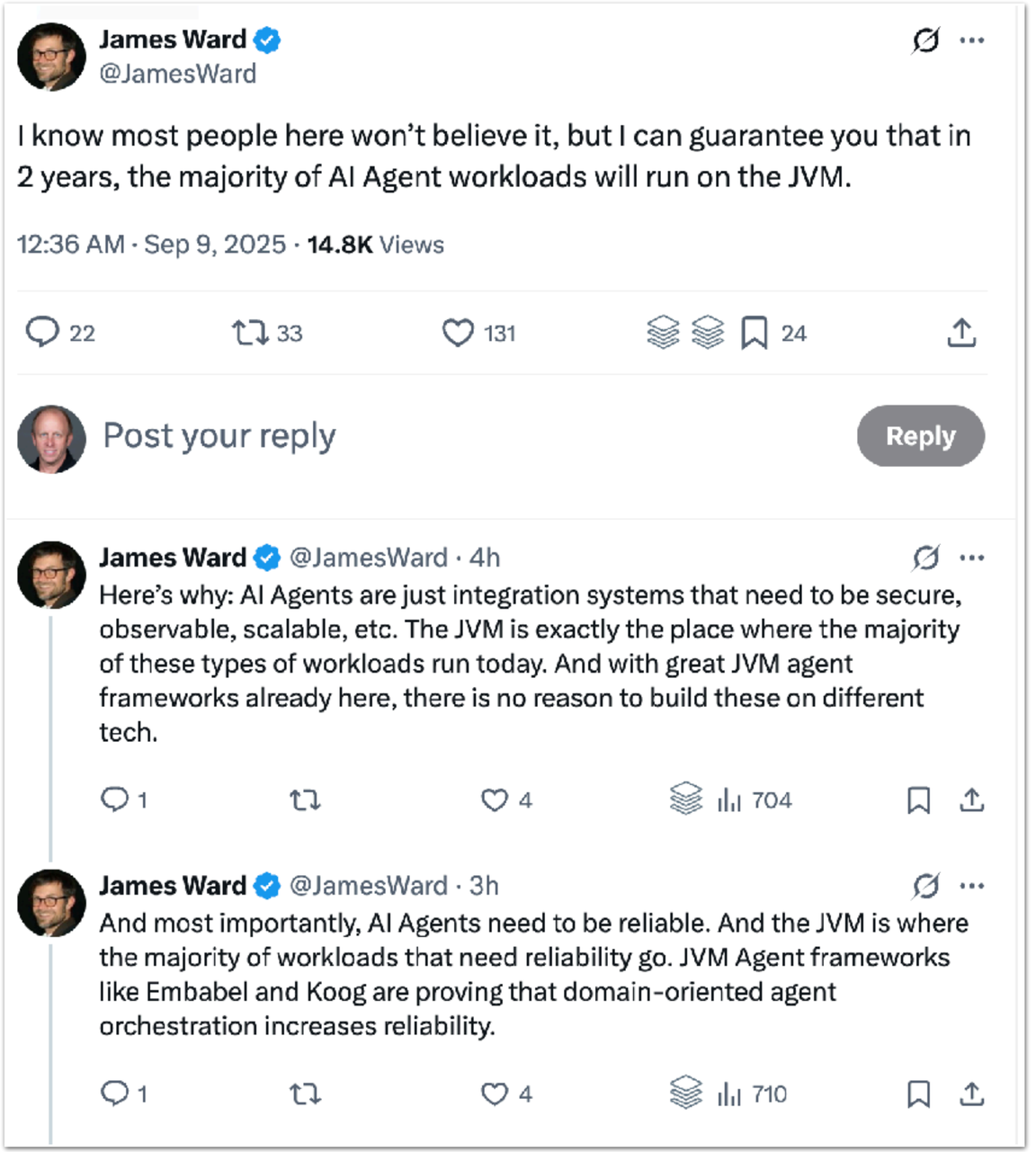
- **Transformers (Hugging Face)** – state-of-the-art NLP (and increasingly multimodal).

THE JAVA AI OPPORTUNITY

- **Enterprise Reality Check** Most large-scale enterprise applications run on the JVM. When these organizations want to integrate AI capabilities, they face a choice: build isolated Python services that require new infrastructure and teams, or leverage their existing **JVM** expertise and infrastructure.
- **Performance and Scalability** The JVM delivers fast, reliable performance for AI applications in production. While Python is great for training models, the JVM shines when serving those models to real users where speed and stability are critical.
- **Mature Ecosystem** The JVM's proven ecosystem including Spring for enterprise apps, Kafka for streaming, and Spark for distributed processing gives AI developers a solid foundation of battle-tested tools rather than building everything from scratch.
- **Language Diversity Advantage** The JVM lets you pick the right language for each AI task. Use Java for enterprise stability, or Kotlin and Groovy for a more Python-like experience, all running on the same platform.
- **Developer Talent Pool** There are millions of JVM developers worldwide. Training existing teams on JVM-based AI tools is often faster and more cost-effective than hiring specialized Python AI teams or forcing Java developers to context-switch between languages.

James Ward

Principal Developer Advocate at AWS



Building and Training
Large Language Models

Consuming / Integration
Large Language Models

```
#!/bin/bash
echo "Calling Open AI..."
MY_OPENAI_KEY=“YOUR_API_KEY_HERE”
PROMPT=“Tell me an interesting fact about Java"

curl https://api.openai.com/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $MY_OPENAI_KEY" \
  -d '{"model": “gpt-4o”, "messages": [{"role":"user", "content": "'"${PROMPT}"'"}] }'
```

```
{
  "id": "chatcmpl-ABNbjZ5oRbo72OevnCX2arPufJCYK",
  "object": "chat.completion",
  "created": 1727275719,
  "model": "gpt-4o-2024-05-13",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "Sure! Did you know that Java was initially designed with interactive television in mind? James Gosling and his team at Sun Microsystems started the project in 1991 under the name \"Oak.\" The name was later changed to \"Java\" after discovering there was already a programming language called Oak. Java's versatility has made it one of the most popular programming languages for a wide range of applications, far beyond its initial intended use for TV set-top boxes!",
        "refusal": null
      },
      "logprobs": null,
      "finish_reason": "stop"
    }
  ],
  "usage": {
    "prompt_tokens": 14,
    "completion_tokens": 90,
    "total_tokens": 104,
    "completion_tokens_details": {
      "reasoning_tokens": 0
    }
  },
  "system_fingerprint": "fp_e375328146"
}
```

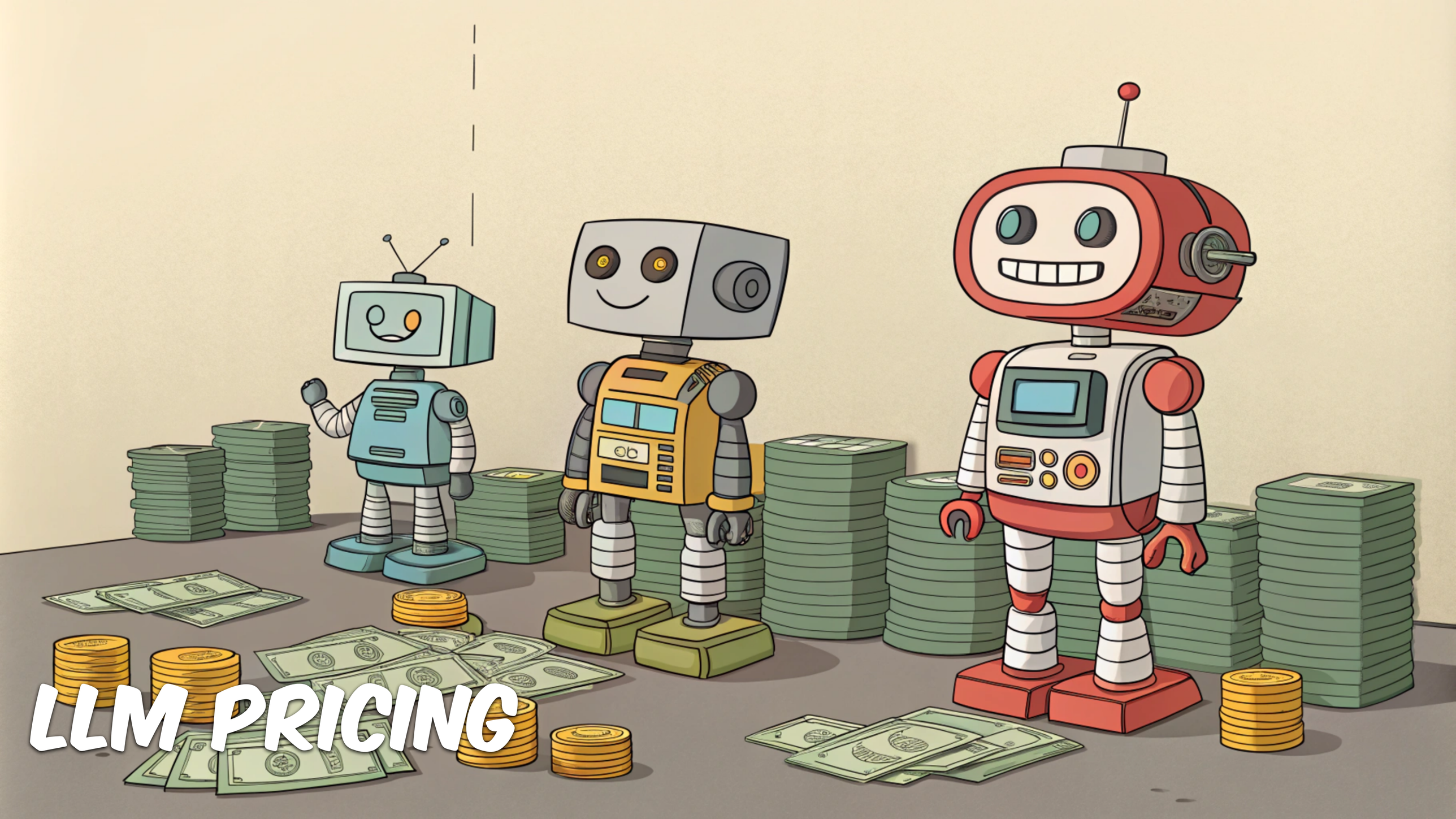
```java
public static void main(String[] args) throws IOException, InterruptedException {
    var apiKey = "YOUR_API_KEY_HERE";
    var body = """
            {
                "model": "gpt-4o",
                "messages": [
                    {
                        "role": "user",
                        "content": "Tell me an interesting fact about Java"
                    }
                ]
            }""";

    HttpRequest request = HttpRequest.newBuilder()
            .uri(URI.create("https://api.openai.com/v1/chat/completions"))
            .header("Content-Type", "application/json")
            .header("Authorization", "Bearer " + apiKey)
            .POST(HttpRequest.BodyPublishers.ofString(body))
            .build();

    var client = HttpClient.newHttpClient();
    var response = client.send(request, HttpResponse.BodyHandlers.ofString());
    System.out.println(response.body());
}
```

WHAT ARE THE CHALLENGES

What does a framework provide you?

- Model abstraction and portability
- Unified API across providers
- Structured output parsing
- Prompt template management
- Token counts & cost management
- Retry logic and error handling
- Response streaming and async processing
- Memory and conversation management
- Embedding and vector operations
- Function calling integration
- Observability and monitoring
- Security and compliance

SPRING AI PROVIDES US SO MUCH MORE THAN A FACILITY FOR MAKING REST API CALLS


LLM PRICING

Free

Explore how AI can help with everyday tasks

- ✓ Access to GPT-4.1 mini
- ✓ Real-time data from the web with search
- ✓ Limited access to GPT-4o, OpenAI o4-mini, and deep research
- ✓ Limited access to file uploads, data analysis, image generation, and voice mode
- ✓ Code edits with the ChatGPT desktop app for macOS
- ✓ Use custom GPTs

Have an existing plan? See billing help

$0 / month

Get Free ↗

Plus

Level up productivity and creativity with expanded access

- ✓ Everything in Free
- ✓ Extended limits on messaging, file uploads, data analysis, and image generation
- ✓ Standard and advanced voice mode with video and screensharing
- ✓ Access to deep research and multiple reasoning models (OpenAI o3, OpenAI o4-mini, and OpenAI o4-mini-high)
- ✓ Access to a research preview of GPT-4.5, our largest model yet, and GPT-4.1, a model optimized for coding tasks
- ✓ Create and use projects, tasks, and custom GPTs
- ✓ Opportunities to test new features

$20 / month

Get Plus ↗ Limits apply >

Pro

Get the best of OpenAI with the highest level of access

- ✓ Everything in Plus
- ✓ Unlimited access to all reasoning models and GPT-4o
- ✓ Unlimited access to advanced voice, with higher limits for video and screensharing
- ✓ Access to OpenAI o1 pro mode, which uses more compute for the best answers to the hardest questions
- ✓ Extended access to deep research
- ✓ Extended access to Sora video generation
- ✓ Access to a research preview of Operator
- ✓ Access to research preview of Codex agent

Unlimited subject to abuse guardrails. Learn more

$200 / month

Get Pro ↗

GPT-5

The best model for coding and agentic tasks across industries

Price

Input:
$1.250 / 1M tokens

Cached input:
$0.125 / 1M tokens

Output:
$10.000 / 1M tokens

GPT-5 mini

A faster, cheaper version of GPT-5 for well-defined tasks

Price

Input:
$0.250 / 1M tokens

Cached input:
$0.025 / 1M tokens

Output:
$2.000 / 1M tokens

GPT-5 nano

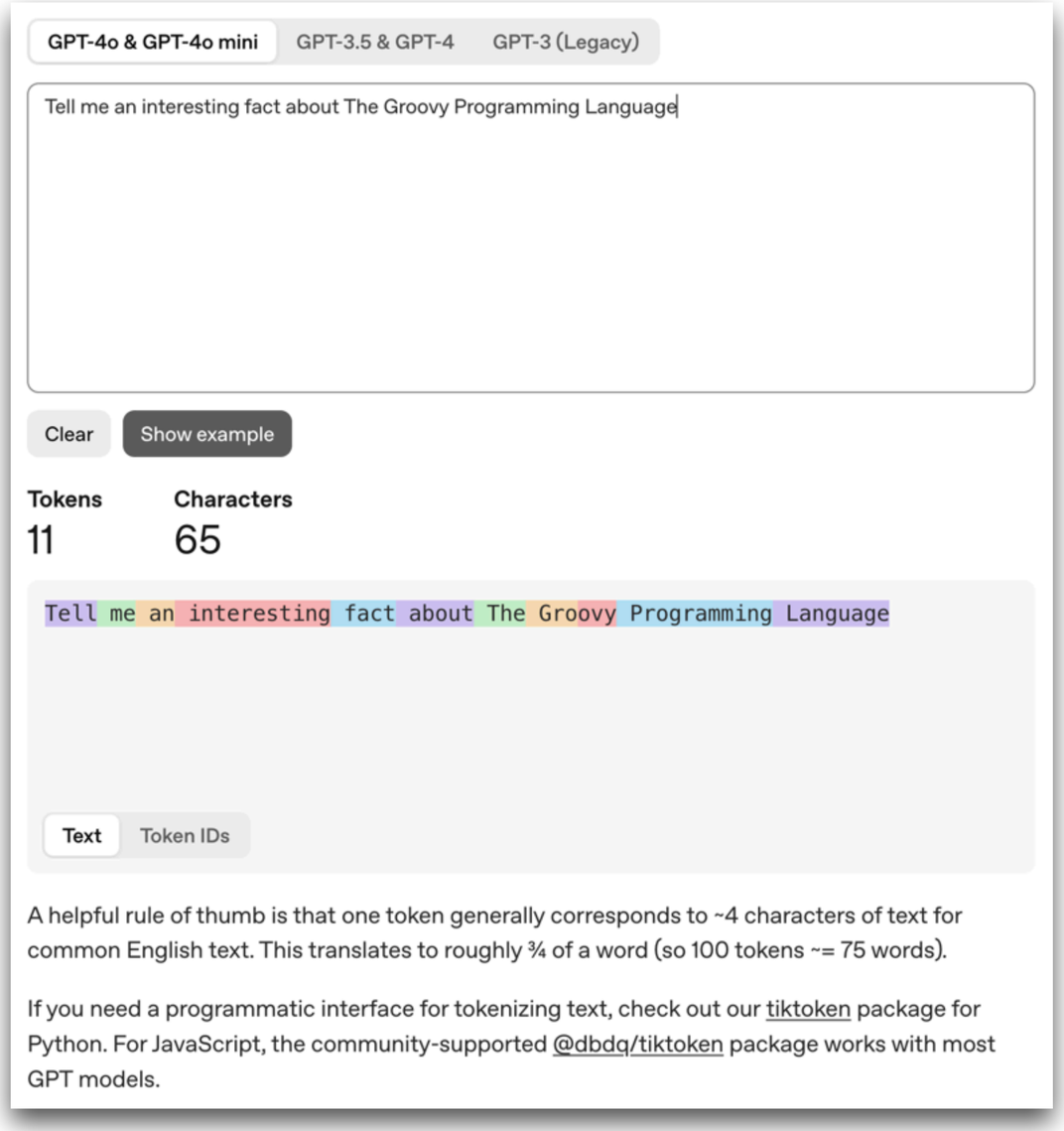
The fastest, cheapest version of GPT-5 —great for summarization and classification tasks

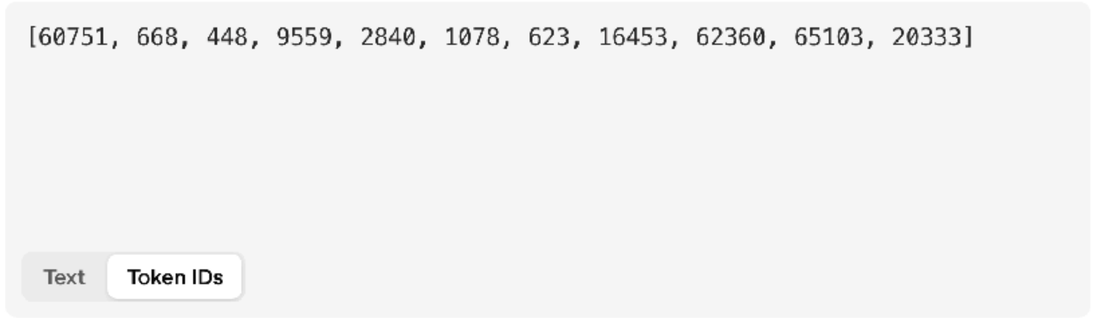
Price

Input:
$0.050 / 1M tokens

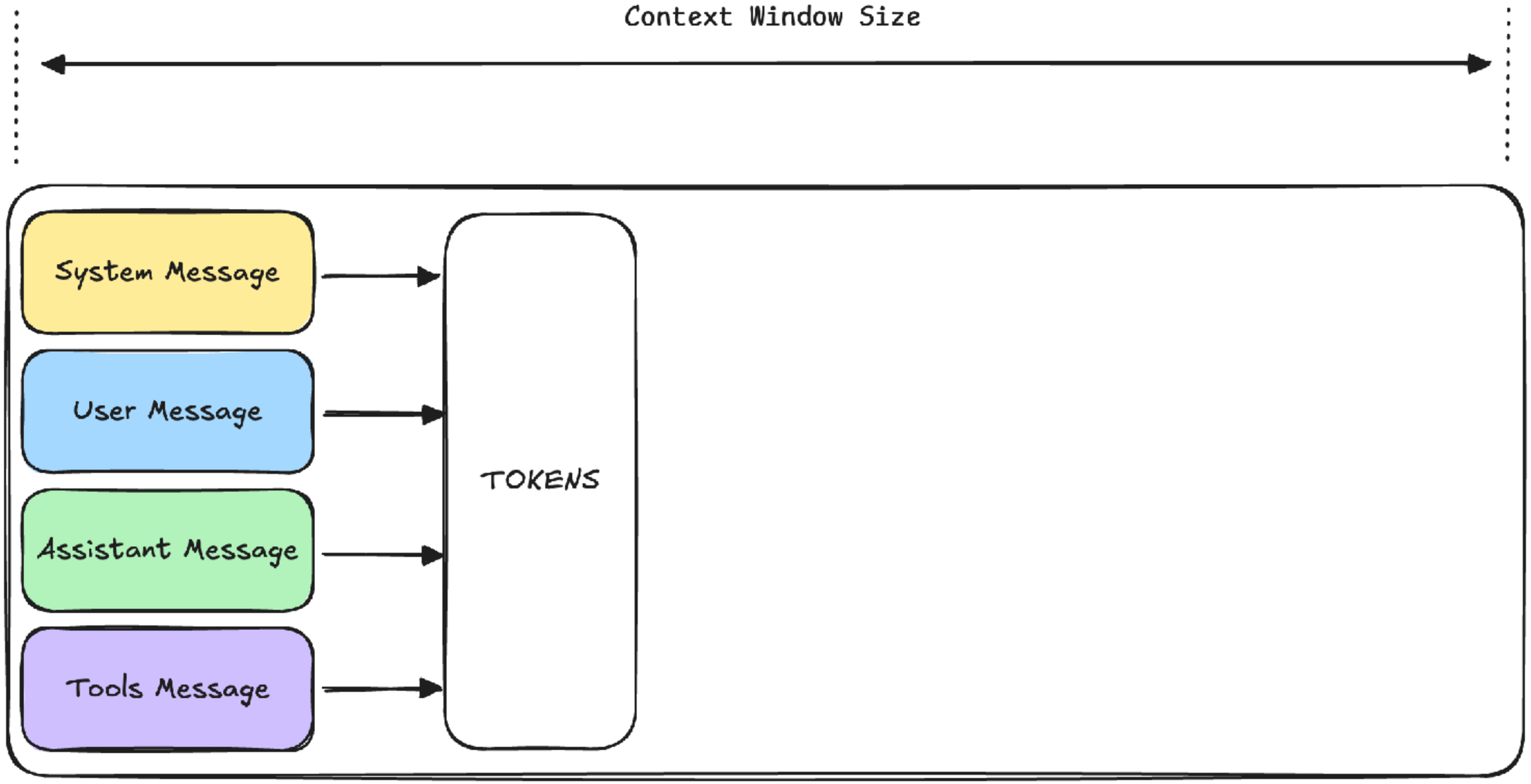
Cached input:
$0.005 / 1M tokens

Output:
$0.400 / 1M tokens


GPT-4o & GPT-4o mini
GPT-3.5 & GPT-4
GPT-3 (Legacy)
Tell me an interesting fact about The Groovy Programming Language
Clear
Show example
Tokens
11
Characters
65
Tell me an interesting fact about The Groovy Programming Language
Text
Token IDs
A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).
If you need a programmatic interface for tokenizing text, check out our tiktoken package for Python. For JavaScript, the community-supported @dbdq/tiktoken package works with most GPT models.


[60751, 668, 448, 9559, 2840, 1078, 623, 16453, 62360, 65103, 20333]
Text
Token IDs

Context Window



Model	Context window	Pricing (per 1M input / output tokens)
GPT-5 (OpenAI)	~400K tokens	$1.25 / $10.00 (cached input: $0.125)
GPT-5 Mini (OpenAI)	~400K tokens	$0.25 / $2.00 (cached input: $0.025)
GPT-5 Nano (OpenAI)	~400K tokens	$0.05 / $0.40 (cached input: $0.005)
Claude Sonnet 4 (Anthropic)	200K tokens	$3.00 / $15.00
Claude Opus 4.1 (Anthropic)	200K tokens (32K output)	$15.00 / $75.00
Gemini 2.5 Flash-Lite (Google)	1M tokens	$0.10 / $0.40
Gemini 2.5 Flash (Google)	1M tokens	$0.30 / $1.25
Gemini 2.5 Pro (Google)	1M tokens (2M roadmap)	≤200K: $1.25 / $10.00 • >200K: $2.50 / $15.00
Grok 3 (xAI)	131K tokens	$3.00 / $15.00 (cached input: $0.75)

SPRING AI INTRODUCTION

GETTING STARTED

A Tour of Spring AI Features

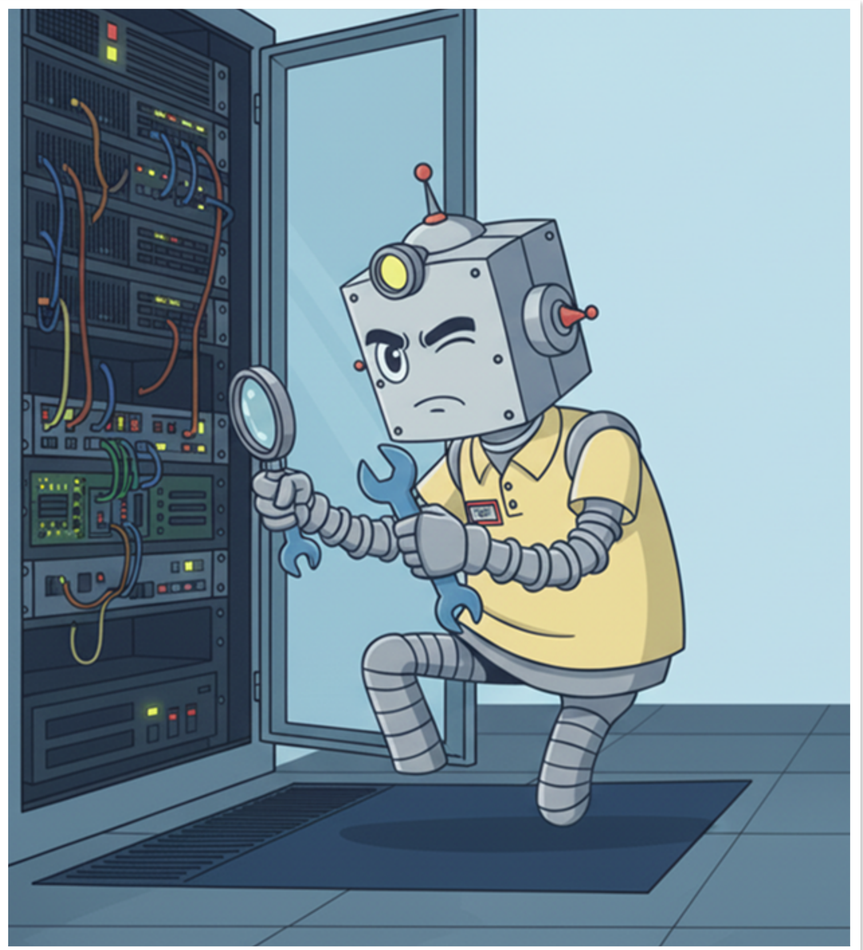
- Spring AI Reference Documentation
  - Currently Version: 1.1.2
- Sign up for an API Key - **OpenAI**
- start.spring.io
- Chat Client & Chat Model
  - Blocking vs Non Blocking (Streaming Responses)
  - Response Types (Content / ChatResponse
- Spring AI Features (Prompt Templates, Structured Output & more...)

CHECK OUT MY DEMO

LLM LIMITATIONS

PROBLEM STATEMENT

How do we augment the limitations of Large Language Models?



Hallucinations

Stale Data

Bias & Safety

Domain Gaps

Context Window

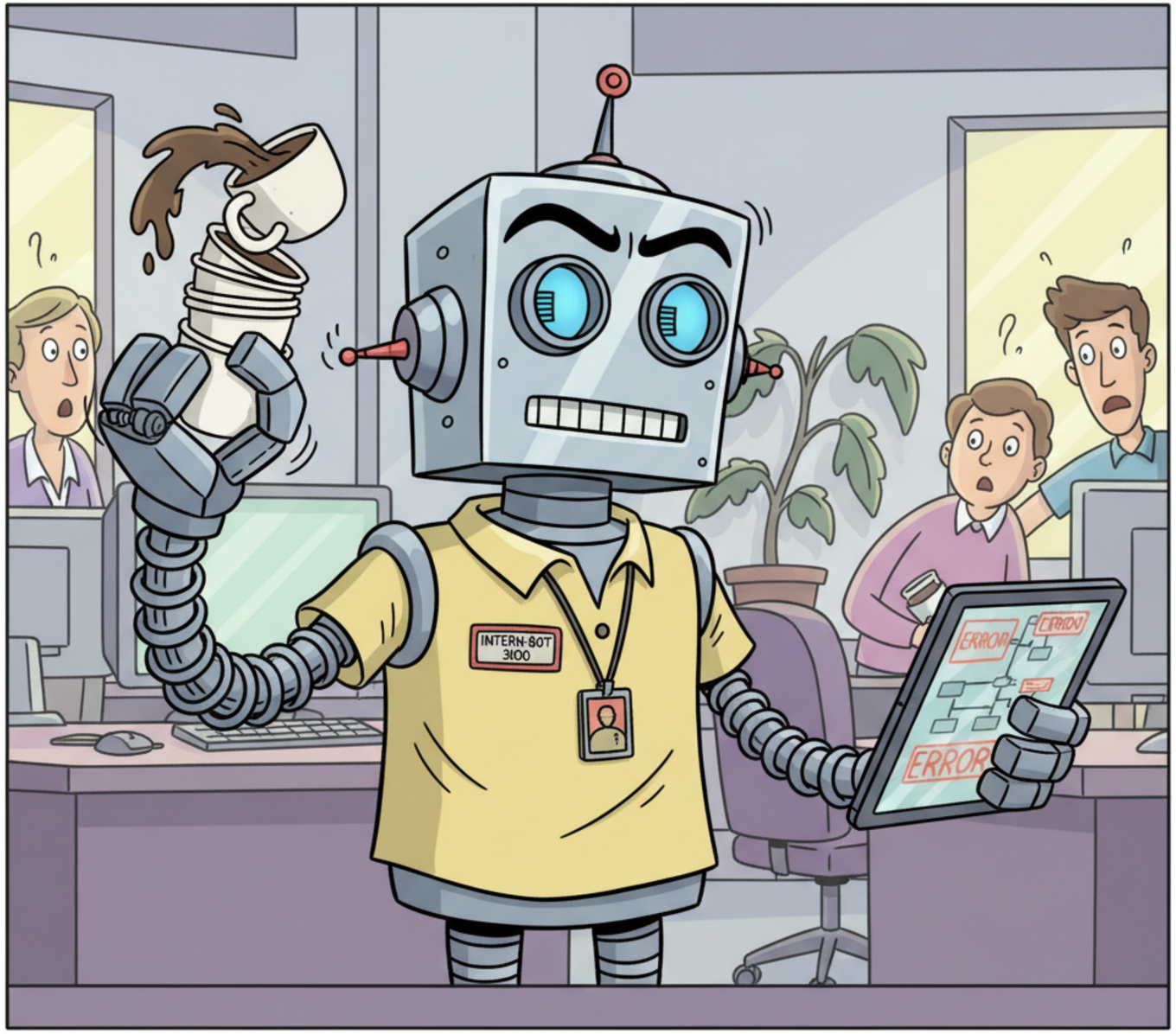
Non-Deterministic

Privacy & Security

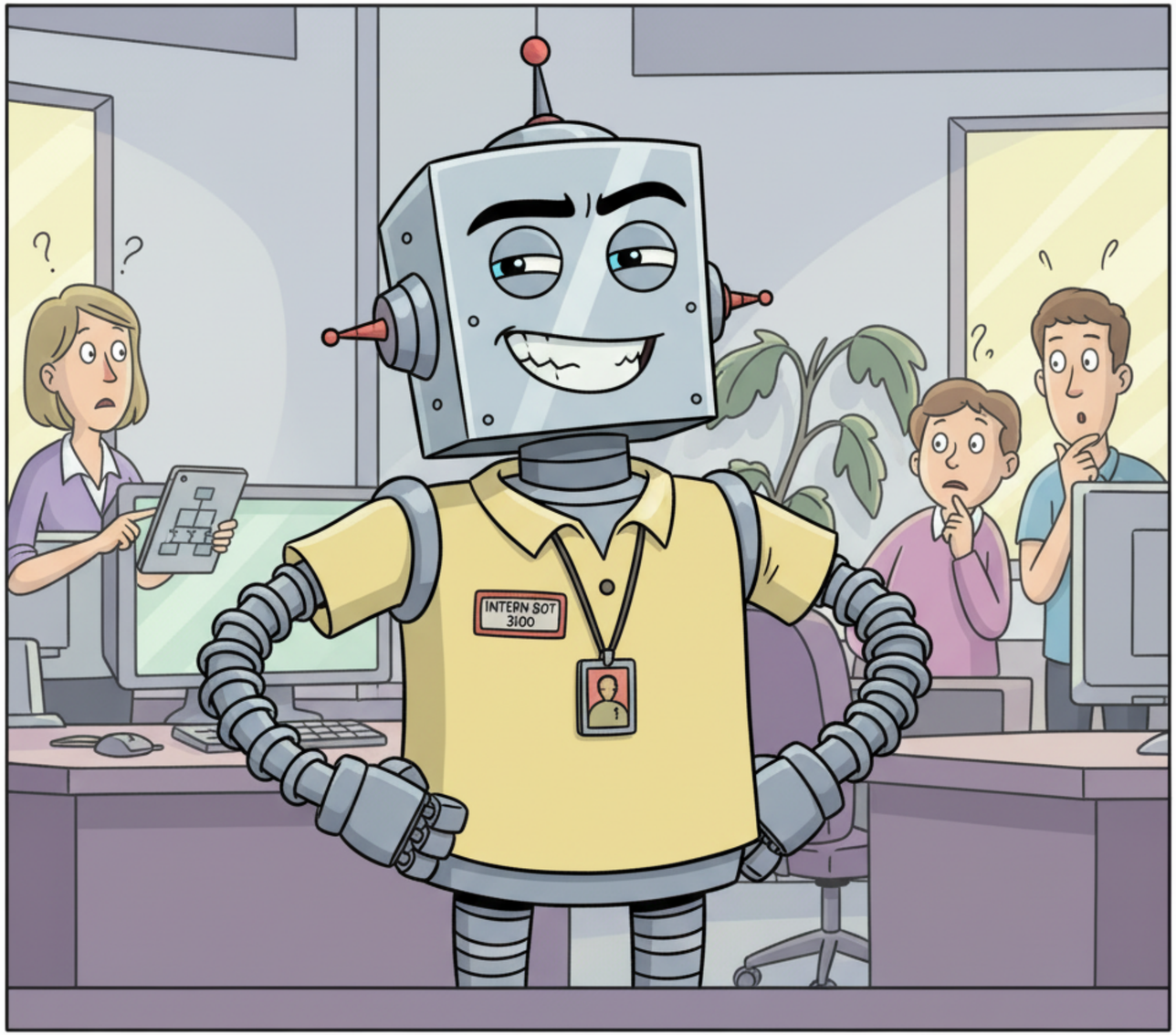
Cost & Latency

Weak Reasoning

LLMS ARE LIKE SUPER-SMART INTERNS



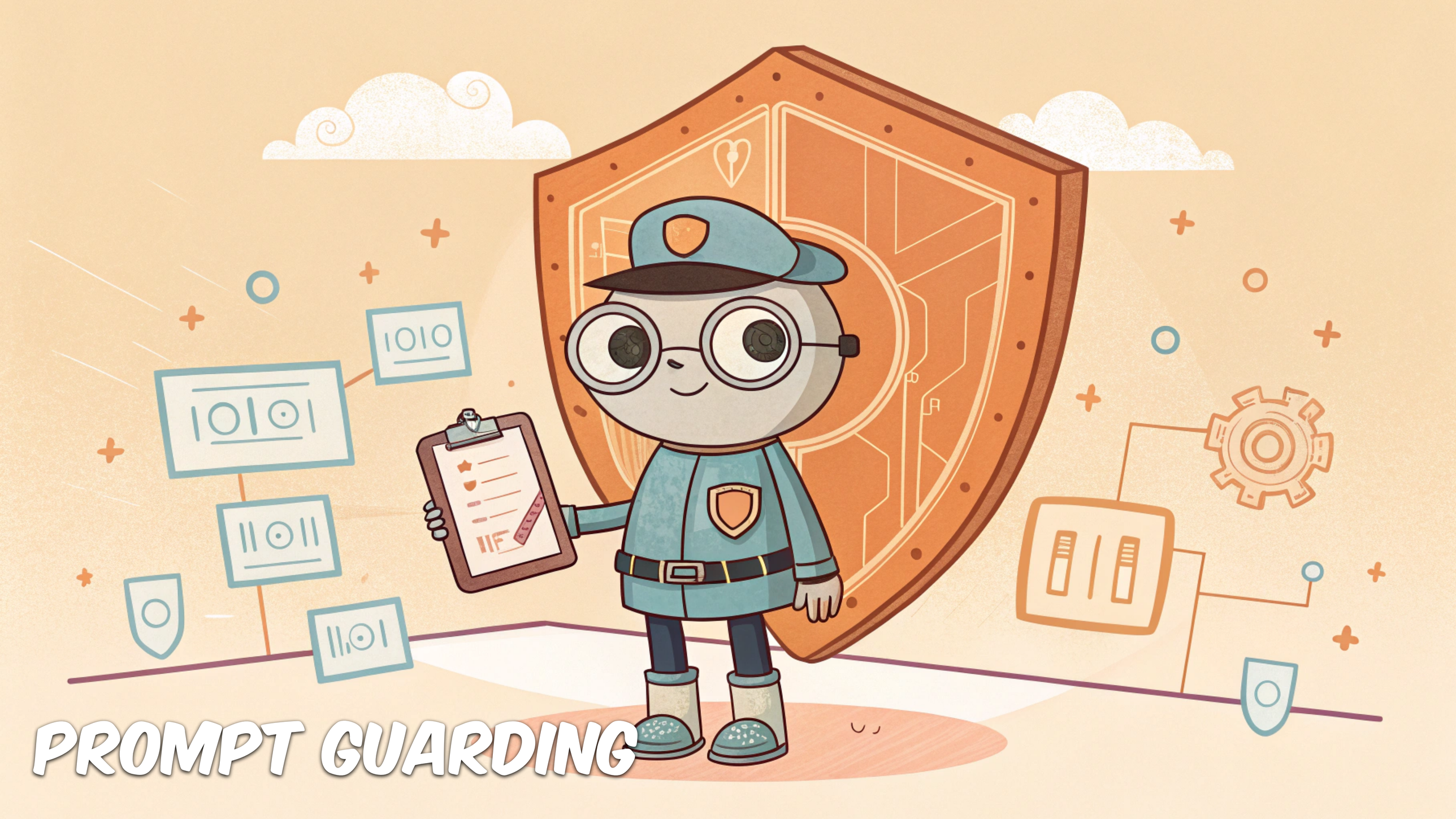
LLMS ARE CONFIDENT

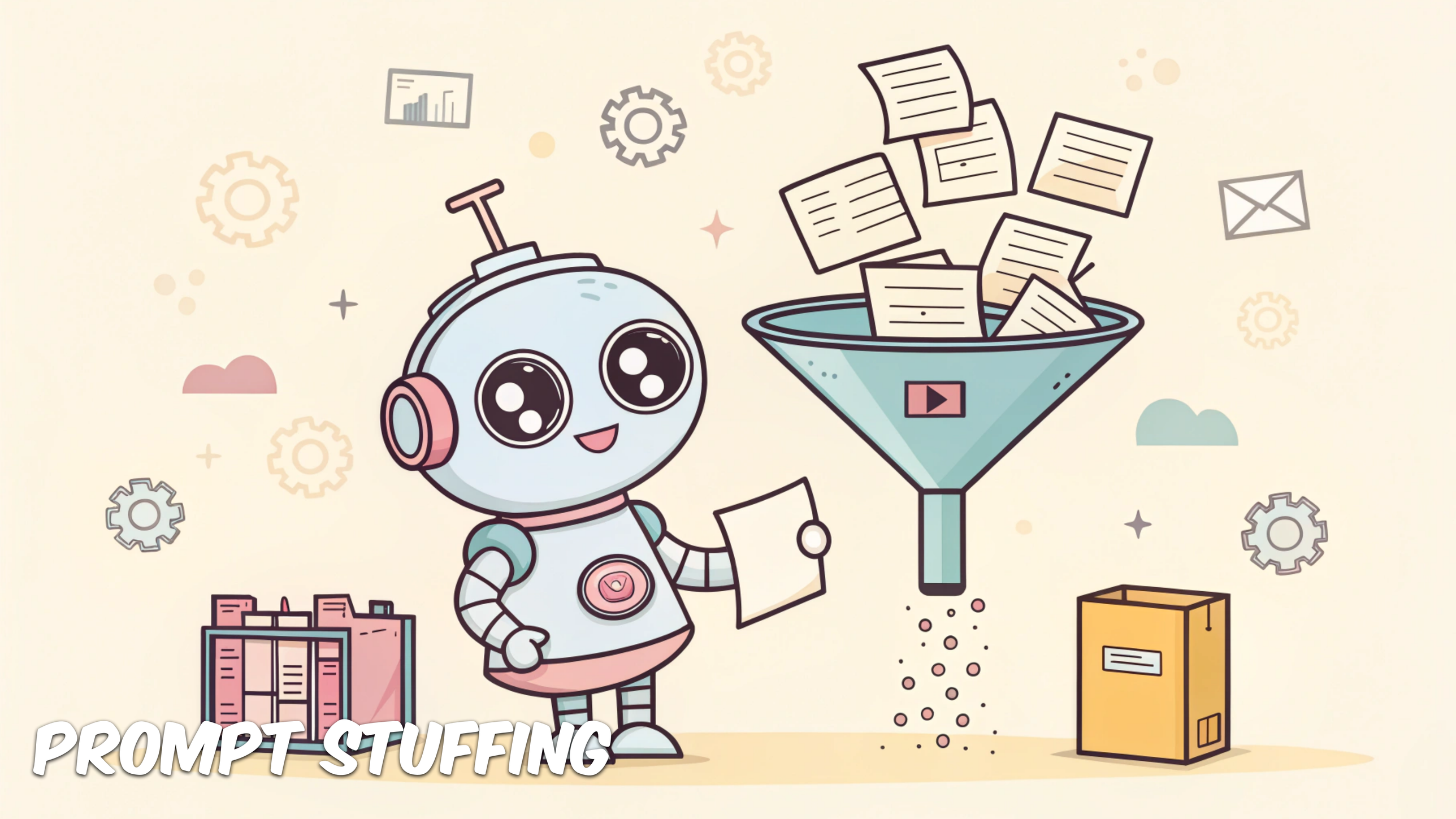


LLM LIMITATIONS

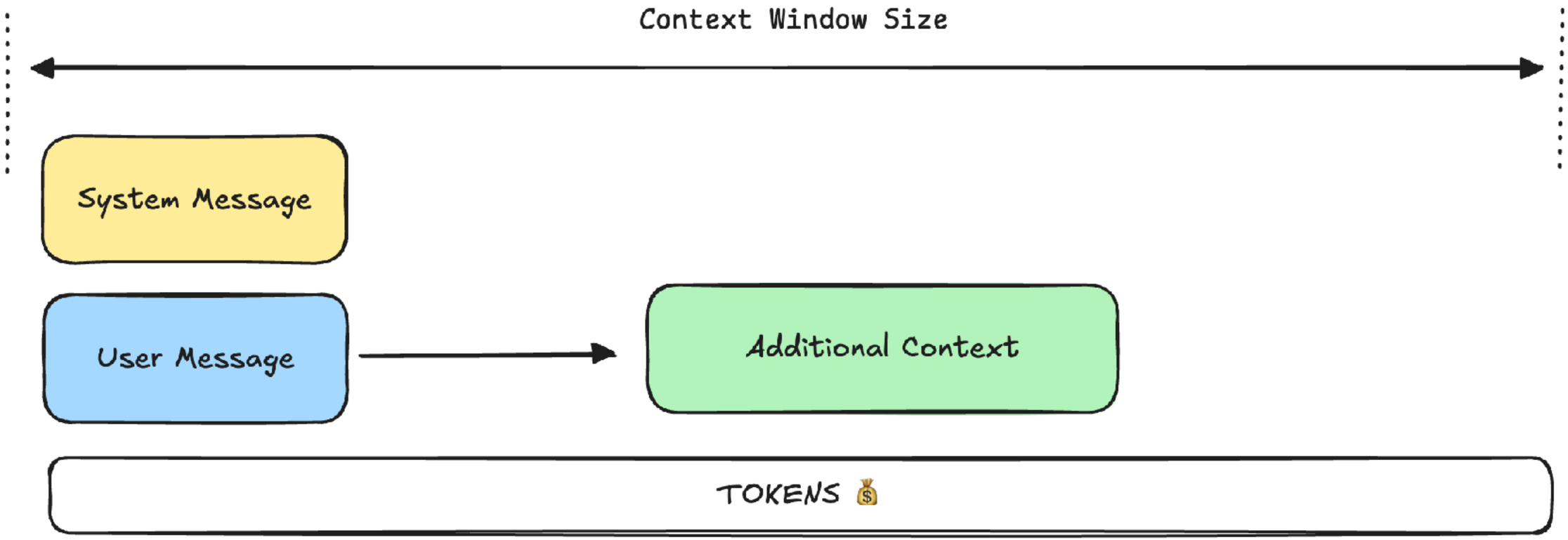
Here's our Swiss-army lineup for taming any limitation

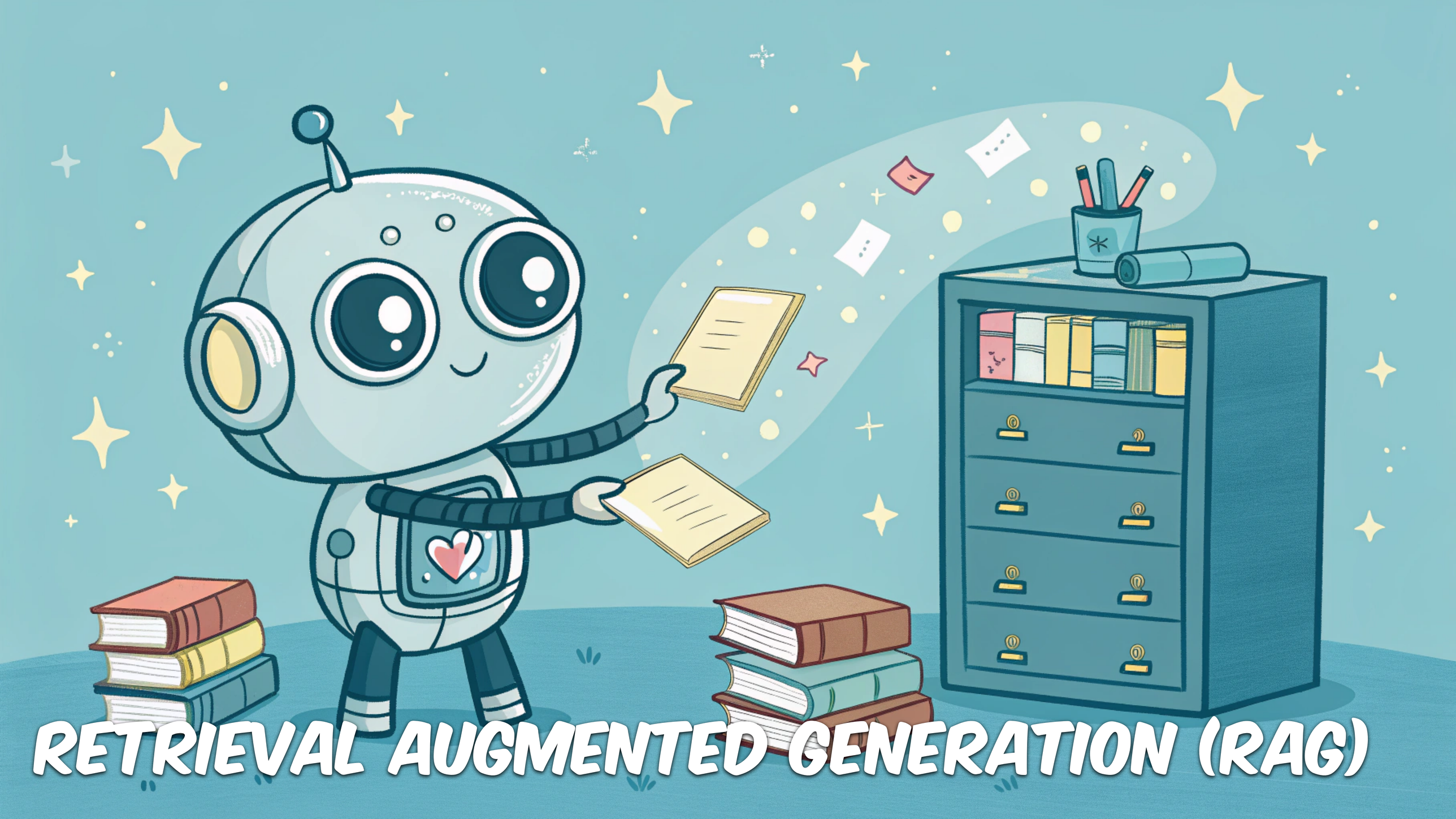
#	Lever	Purpose
1	🛡️ **Prompt Guarding**	Encode rules that constrain the model's behavior (tone, honesty, refusal policy).
2	📄 **Prompt Stuffing / RAG**	Inject fresh, task-specific context so the model quotes facts instead of guessing.
3	🔧 **Tools / Function Calling**	Let the model invoke code or APIs for real-time data, calculations, or business logic.
4	🌐 **MCP (Resources + Tools)**	Package those tools as reusable, versioned endpoints every client can share.


PROMPT GUARDING

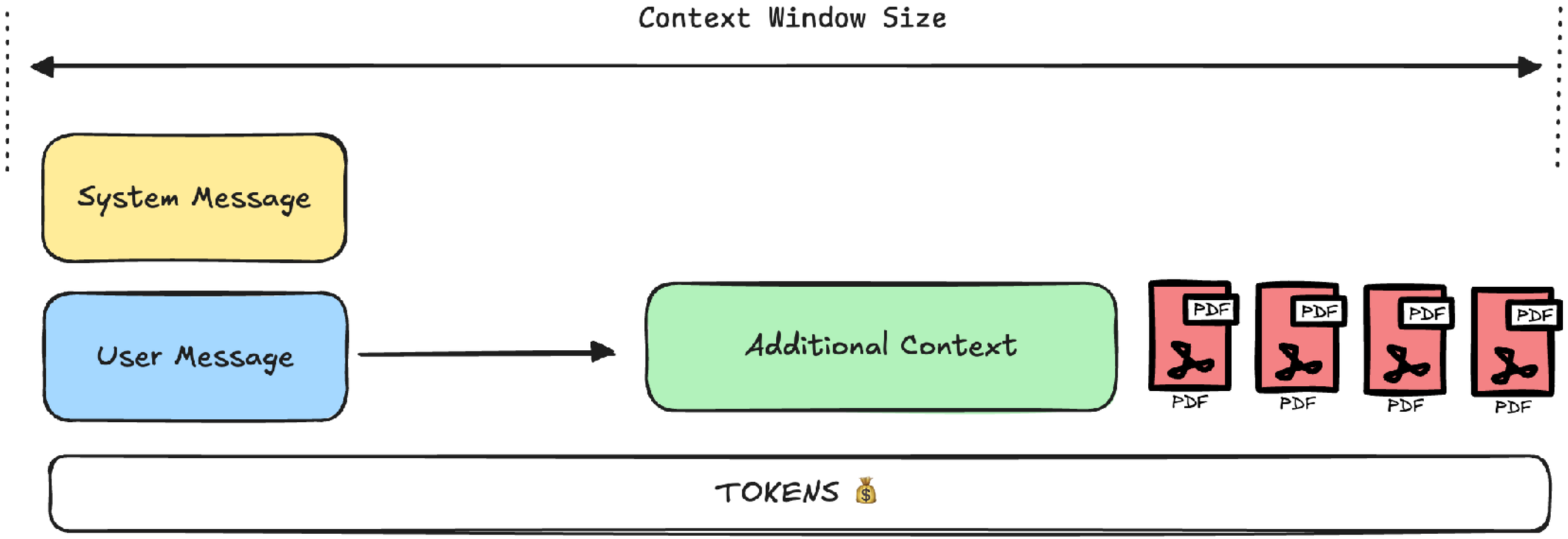

PROMPT STUFFING

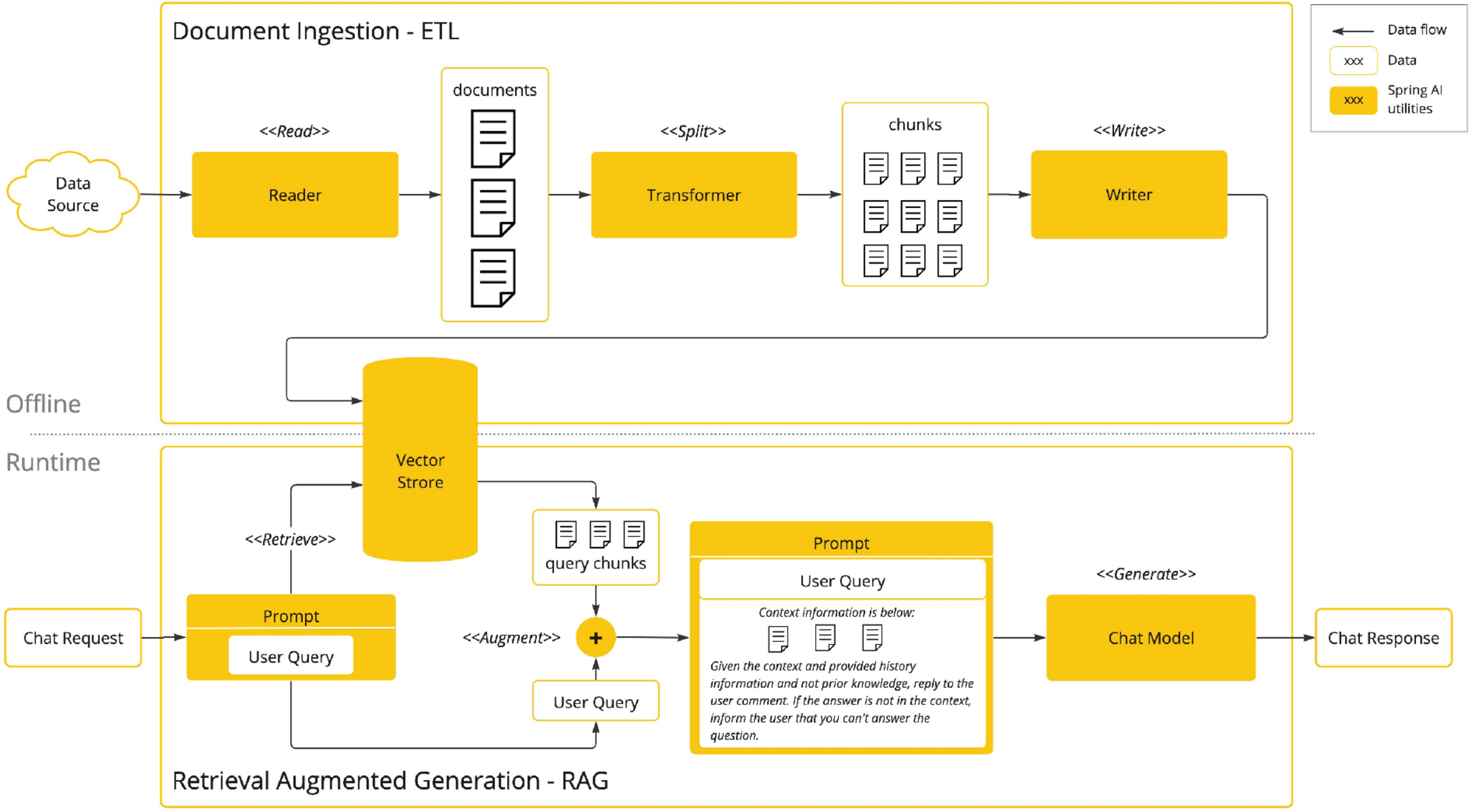
Context Window

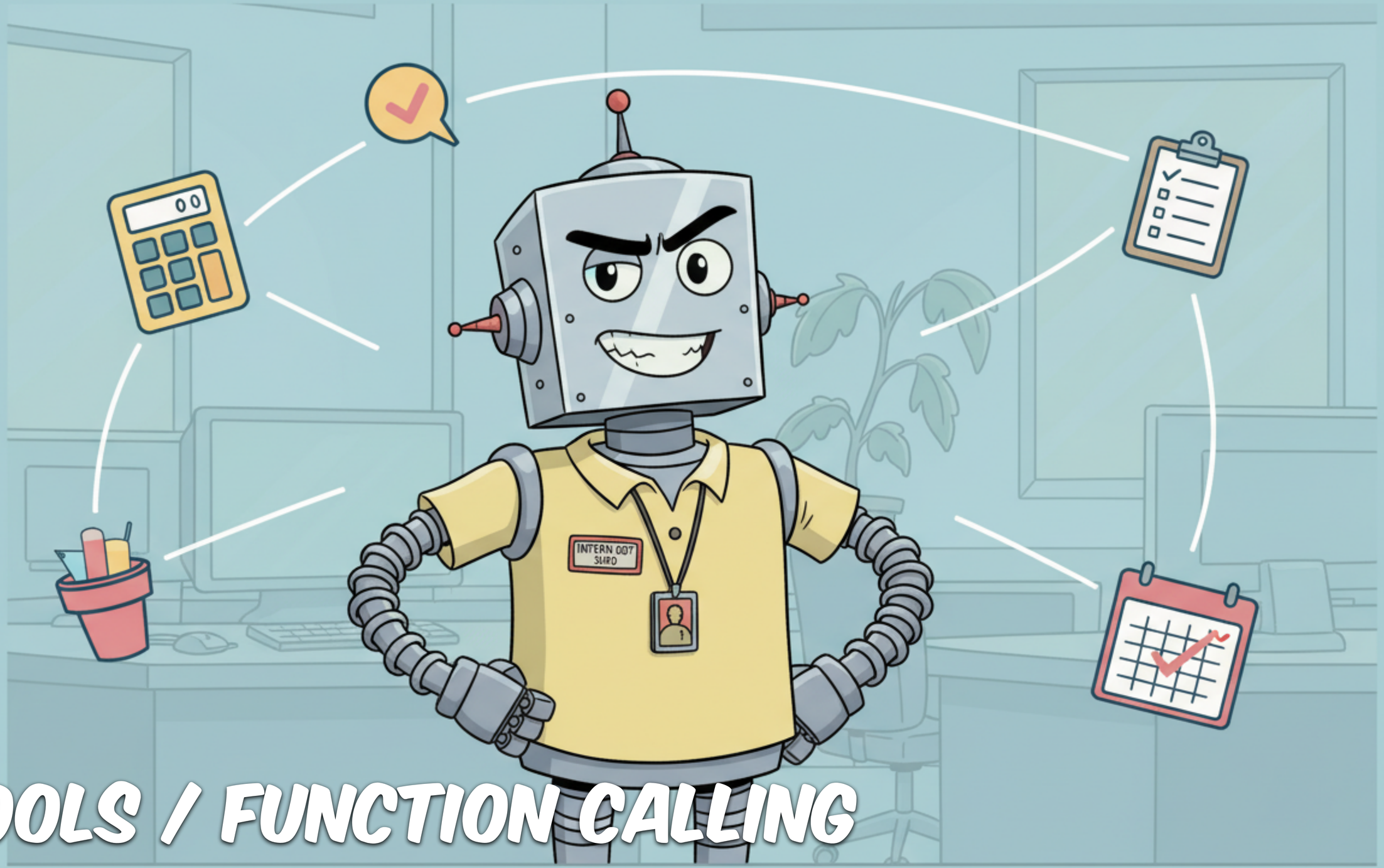



RETRIEVAL AUGMENTED GENERATION (RAG)

Context Window

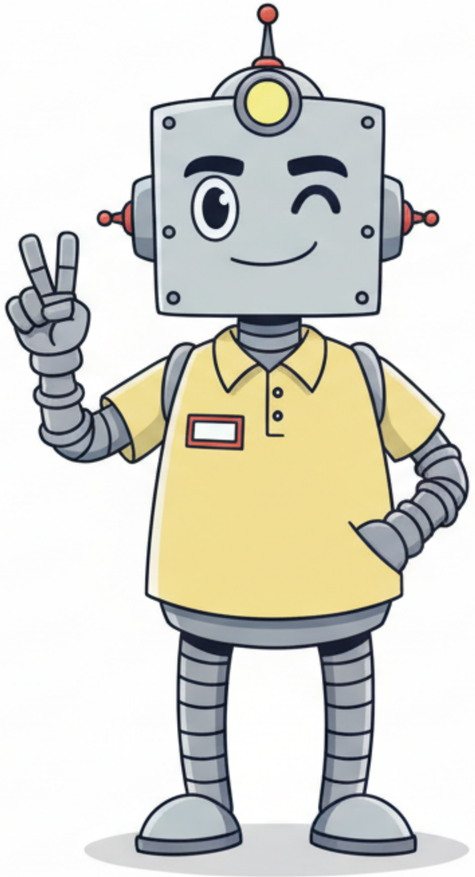



Document Ingestion - ETL
Data flow
xxx
Data
xxx
Spring AI utilities
Data Source
<<Read>>
Reader
documents
<<Split>>
Transformer
chunks
<<Write>>
Writer
Offline
Runtime
Vector Strore
<<Retrieve>>
query chunks
Prompt
User Query
Chat Request
<<Augment>>
+
User Query
Prompt
User Query
Context information is below:
Given the context and provided history information and not prior knowledge, reply to the user comment. If the answer is not in the context, inform the user that you can't answer the question.
<<Generate>>
Chat Model
Chat Response
Retrieval Augmented Generation - RAG



TOOLS / FUNCTION CALLING

Information Retrieval

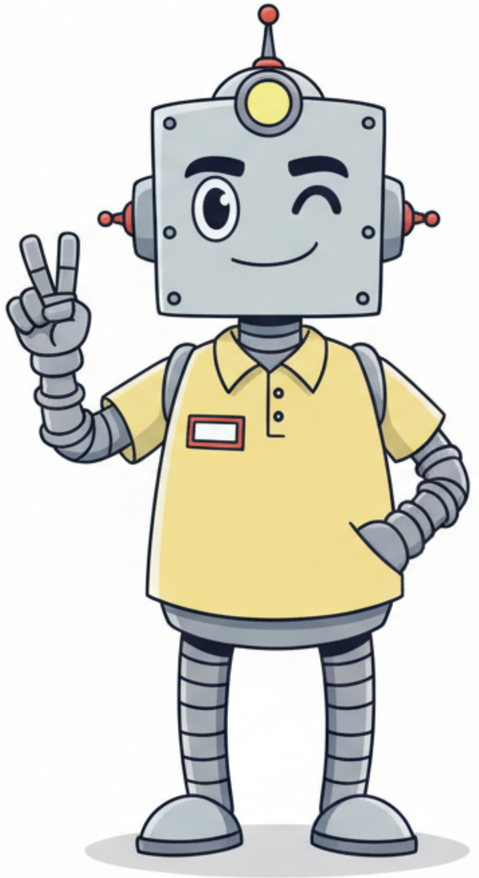


Taking Action

TOOL CALLING / FUNCTION CALLING

Information Retrieval

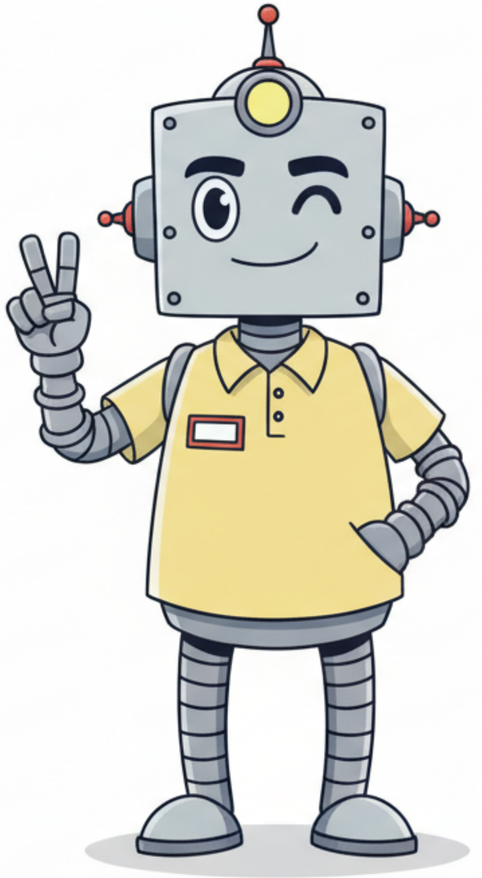
Tools in this category can be used to retrieve information from external sources, such as a database, a web service, a file system, or a web search engine. The goal is to augment the knowledge of the model, allowing it to answer questions that it would not be able to answer otherwise. As such, they can be used in Retrieval Augmented Generation (RAG) scenarios. For example, a tool can be used to retrieve the current weather for a given location, to retrieve the latest news articles, or to query a database for a specific record.



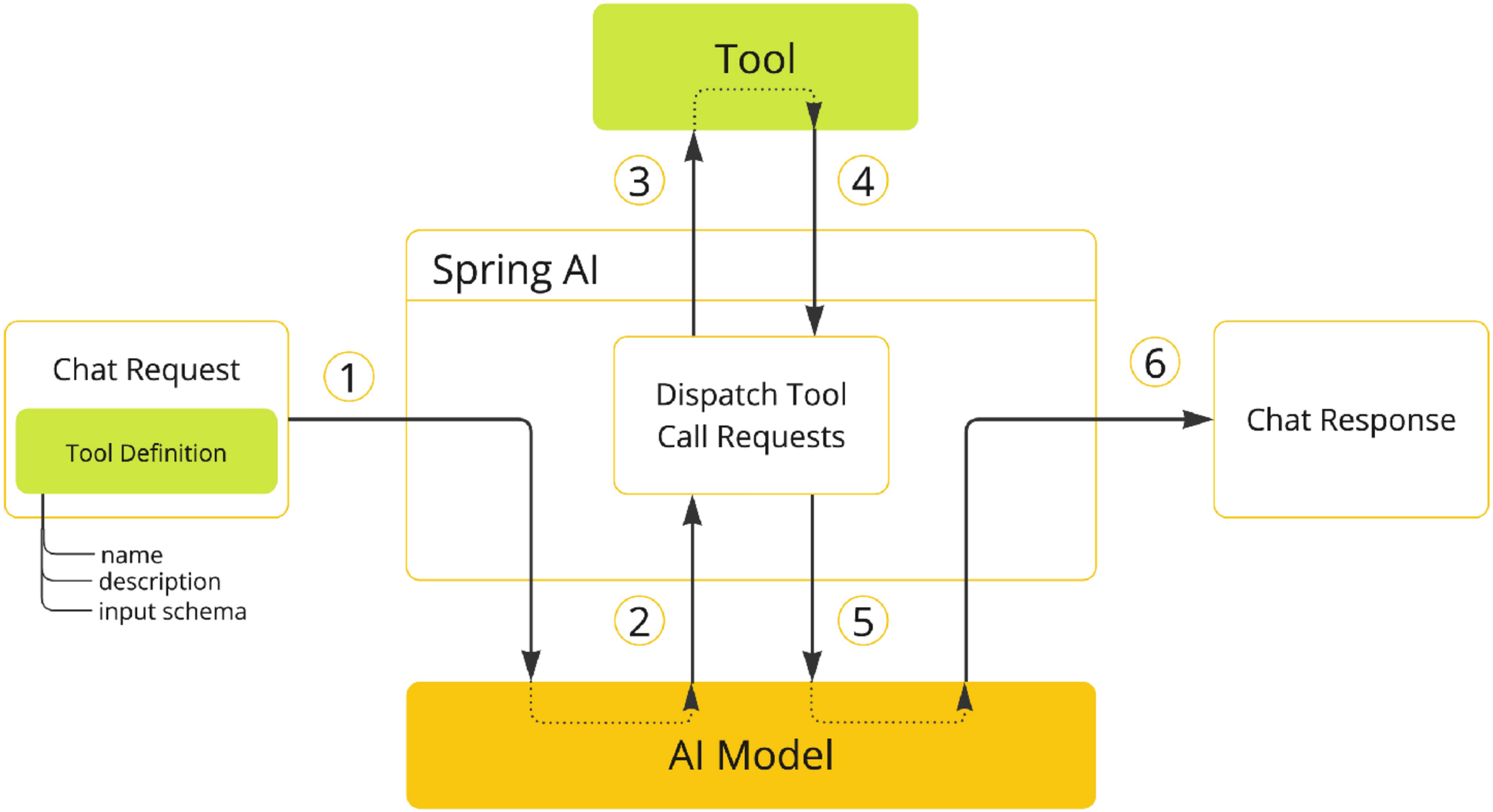
TOOL CALLING / FUNCTION CALLING

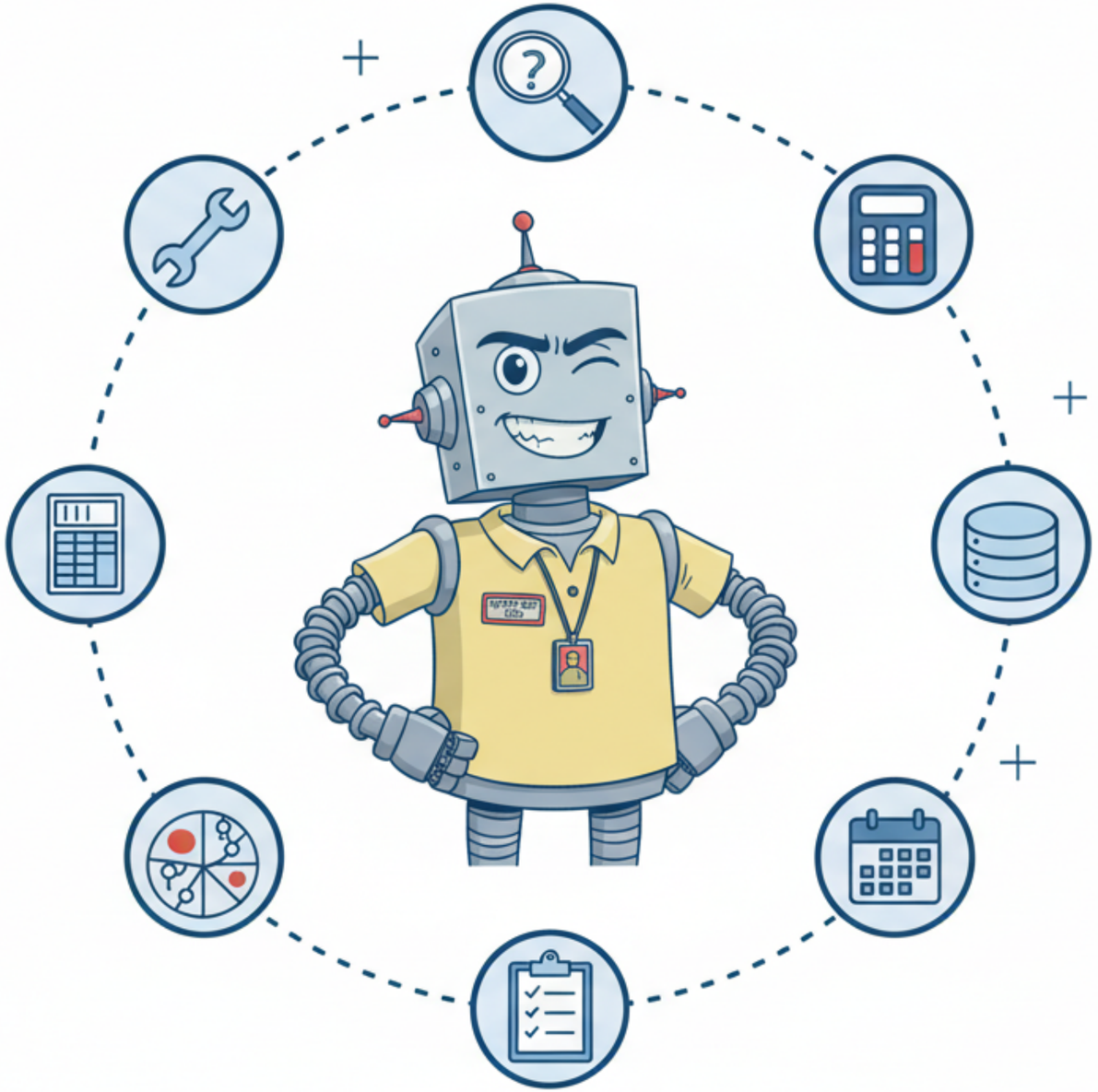
Taking Action

Tools in this category can be used to take action in a software system, such as sending an email, creating a new record in a database, submitting a form, or triggering a workflow. The goal is to automate tasks that would otherwise require human intervention or explicit programming. For example, a tool can be used to book a flight for a customer interacting with a chatbot, to fill out a form on a web page, or to implement a Java class based on an automated test (TDD) in a code generation scenario.



```java
public class DatTimeTools {

    @Tool(description = "Get the current date and time in the user's timezone")
    String getCurrentDateTime() {
        return LocalDateTime.now().atZone(LocaleContextHolder.getTimeZone().toZoneId()).toString();
    }

}
```


Tool
3
4
Spring AI
Chat Request
Tool Definition
name
description
input schema
1
Dispatch Tool
Call Requests
6
Chat Response
2
5
AI Model



CHECK OUT MY DEMO

MCP

How many java sessions are at codemash 2026?

9 steps

Failed to fetch https://codemash.org/session-list/?tag=Java

Tallied Java sessions from fetched list, acknowledged potential incompleteness.

Based on my search of the CodeMash 2026 session list, I found **2 active sessions explicitly tagged with "Java":**

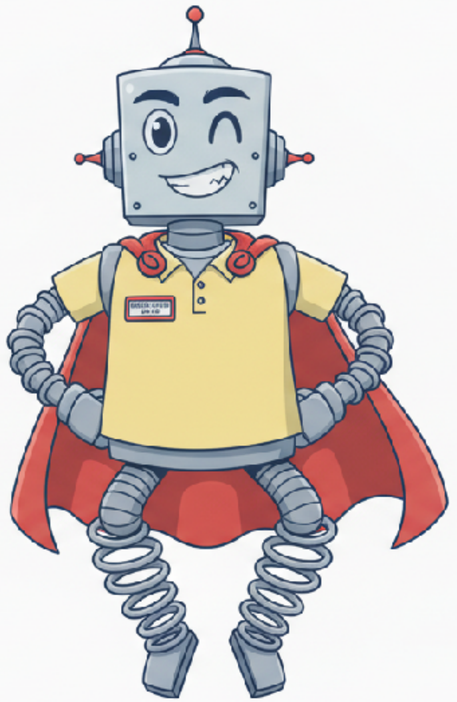
1. **"Demystifying GenAI: Build a ChatGPT App with Vector Store"** - A PreCompiler session presented by Mary Grygleski and Yasmin Rodriguez, tagged with Java, Python, Backend, Machine Learning, Architecture, and System Design
2. **"Creating Intuitive APIs With The Builder Design Pattern"** - A General Session presented by Kelly Morrison, tagged with Java, Object Oriented Programming, and Programming Principles

There was also one additional Java session that has been **cancelled:**

- "(CANCELLED) Scaling from 1 to Infinity: Building Simpler Distributed Applications w/ the Actor Model"

So the total count of active Java sessions is pretty slim this year at CodeMash - just 2. Given your focus on Java and Spring content, this might be a good opportunity if you're looking to submit for future events! The conference seems heavier on JavaScript, .NET/C#, and

MODEL CONTEXT PROTOCOL



WHAT IS MODEL CONTEXT PROTOCOL (MCP)?

MCP is a protocol that defines how to talk to AI models in a consistent, structured way.

MOTIVATION

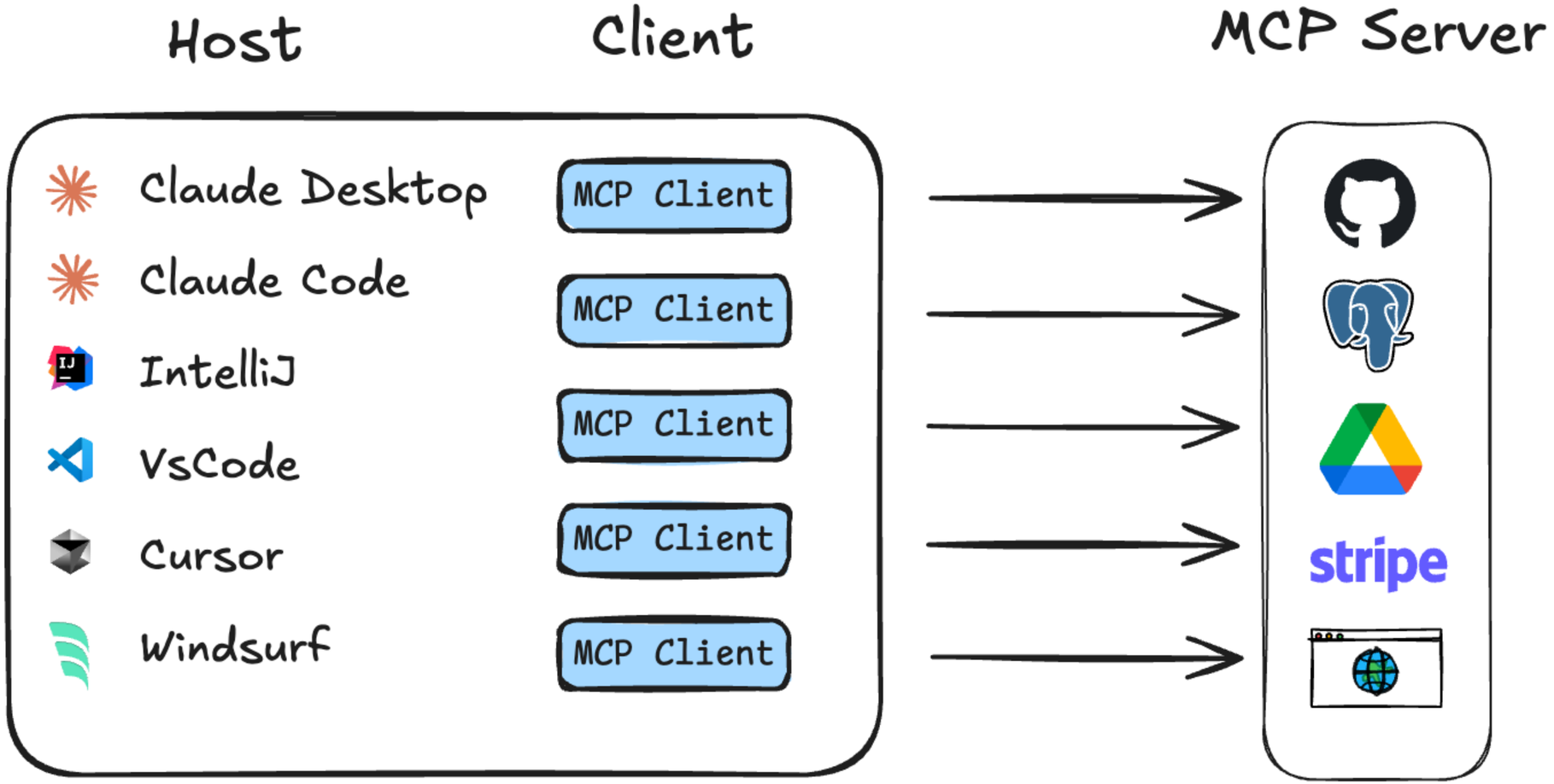
Models are only as good as the **context** you provide them

BENEFITS OF MCP

- **Modularity:** 🧠 This lets the AI app stay light while still being deeply aware of your environment.
- **Reusability:** 💡 Once you write an MCP server (say, one that surfaces project files), it can serve any app that supports the MCP protocol.
- **Language Agnostic:** 👨🏻‍💻 Works across programming languages and environments without language-specific dependencies.
- **Fine-Grained Control:** 🔐 It gives you programmable control over what your AI sees and can do
- **Privacy and Security:** 🛡️ You stay in control of what your AI knows.

COMPONENTS OF MCP

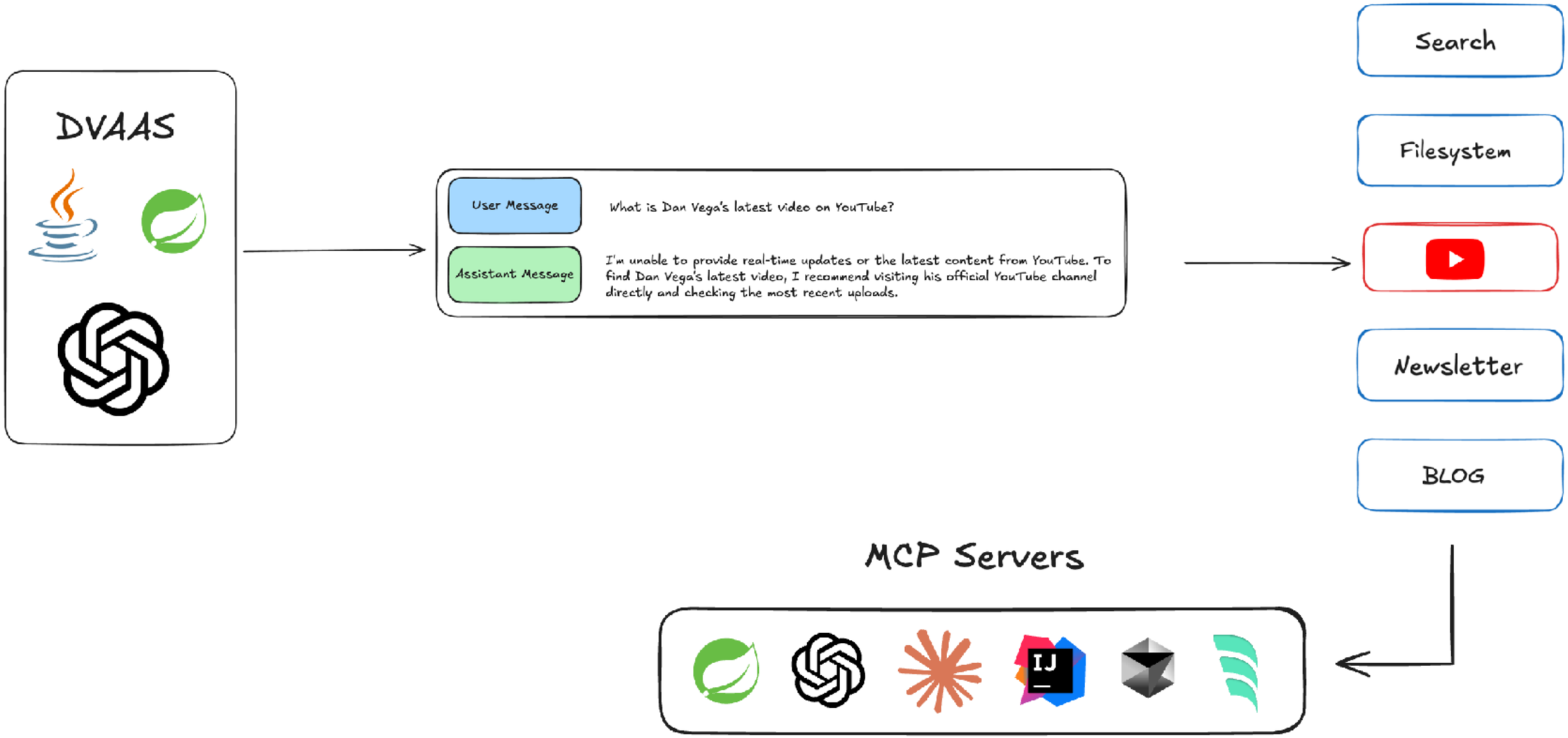
Participants



HOW IS AN MCP SERVER DIFFERENT THAN AN API?

Provider Agnostic

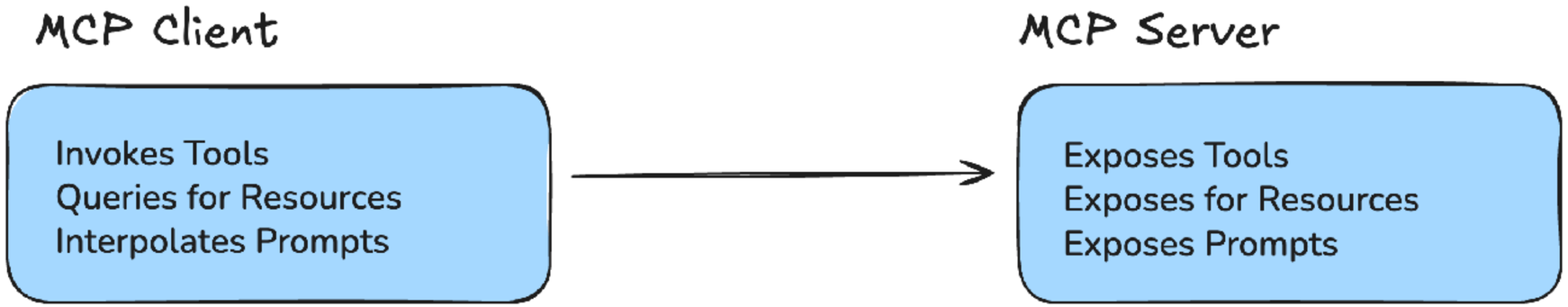
HOW IS THIS DIFFERENT FROM TOOL USE?



PRIMITIVES

MCP PRIMITIVES

MCP primitives are the most important concept within MCP. They define what clients and servers can offer each other. These primitives specify the types of contextual information that can be shared with AI applications and the range of actions that can be performed.



MCP PRIMITIVES: TOOLS

What: Executable functions that AI applications can invoke to perform actions (e.g., file operations, API calls, database queries)

When: AI needs to take action beyond just generating text - when it needs to DO something in the real world

Examples:

- read_file() - Read contents of a specific file
- write_file() - Create or modify files
- execute_sql() - Run database queries
- send_email() - Send messages via email API
- git_commit() - Commit changes to repository
- slack_post() - Send messages to Slack channels
- web_search() - Search the internet
- calculate() - Perform mathematical operations

MCP PRIMITIVES: RESOURCES

What: Data sources that provide contextual information to AI applications (e.g., file contents, database records, API responses)

When: AI needs to understand or reference existing information before responding or taking action

Examples:

- file://project/README.md - Documentation and project files
- database://users/profile/123 - User records and data
- git://repo/commit/history - Version control information
- slack://channel/messages - Chat history and conversations
- calendar://events/today - Schedule and meeting data
- email://inbox/recent - Email content and metadata

MCP PRIMITIVES: PROMPTS

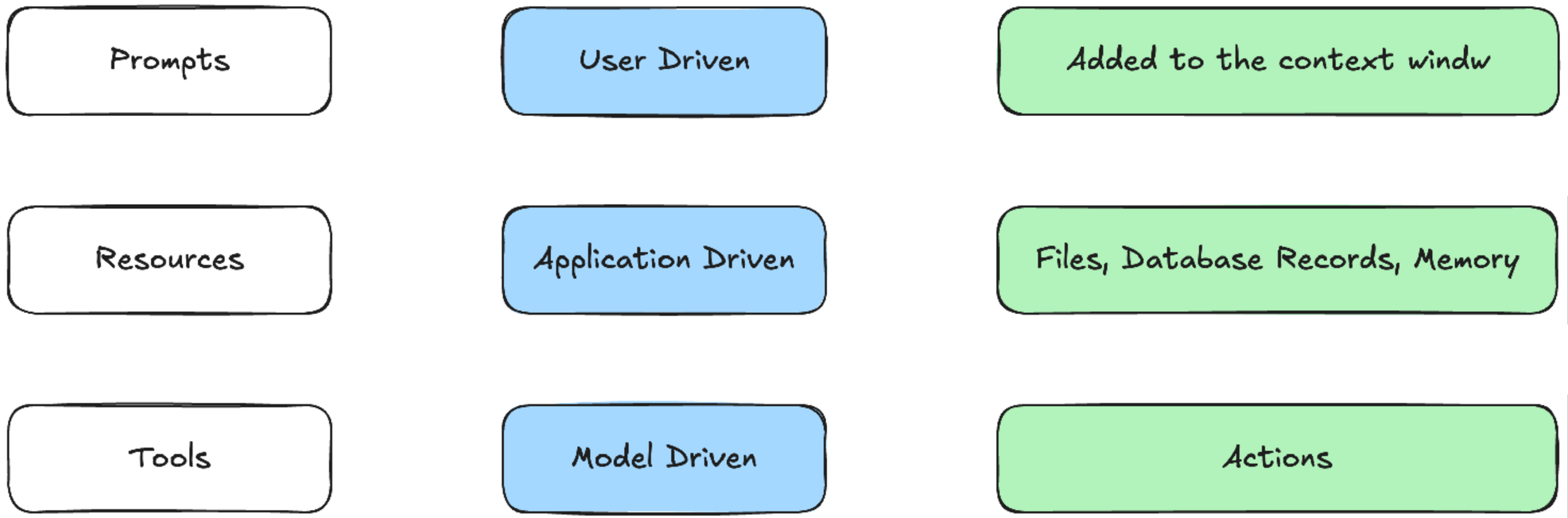
What: Reusable templates that help structure interactions with language models (e.g., system prompts, few-shot examples)

When: You need consistent, well-crafted prompts across different conversations or want to standardize AI behavior patterns

Examples:

- code_reviewer - Template for reviewing pull requests with specific criteria
- meeting_summarizer - Structured format for extracting action items from transcripts
- technical_writer - Guidelines for creating documentation in company style
- bug_triager - Template for categorizing and prioritizing issues
- customer_support - Consistent tone and approach for user interactions
- data_analyst - Framework for interpreting charts and metrics
- project_planner - Structure for breaking down tasks and timelines
- content_moderator - Guidelines for evaluating user-generated content

MCP PRIMITIVES: INTERACTION MODEL



TRANSPORTS

Transports in the Model Context Protocol (MCP) provide the foundation for communication between clients and servers. A transport handles the underlying mechanics of how messages are sent and received.

TRANSPORTS

Standard Input/Output (stdio)

- Building command-line tools
- Implementing local integrations
- Needing simple process communication
- Working with shell scripts

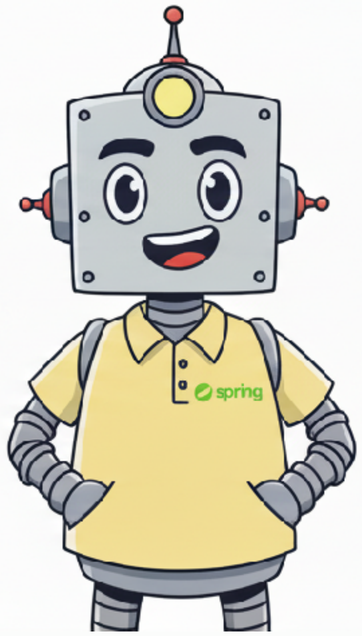
Server-Sent Events (SSE)

- Only server-to-client streaming is needed
- Working with restricted networks
- Implementing simple updates

Streamable HTTP

- Building web-based integrations
- Implementing bidirectional streaming
- Request / Response Streaming
- Working with modern HTTP infrastructure

BUILDING MCP SERVERS IN JAVA / SPRING



SDKs

Copy page

Official SDKs for building with the Model Context Protocol

Build MCP servers and clients using our official SDKs. Choose the SDK that matches your technology stack - all SDKs provide the same core functionality and full protocol support.

Available SDKs

- TypeScript
- Python
- Go
- Kotlin
- Swift
- Java
- C#
- Ruby
- Rust

spring initializr

Project

- ○ Gradle - Groovy
- ○ Gradle - Kotlin
- ● Maven

Language

- ● Java
- ○ Kotlin
- ○ Groovy

Spring Boot

- ○ 4.0.0 (SNAPSHOT)
- ○ 4.0.0 (M1)
- ○ 3.5.5 (SNAPSHOT)
- ● 3.5.4
- ○ 3.4.9 (SNAPSHOT)
- ○ 3.4.8

Project Metadata

Group	dev.danvega
Artifact	kcdc-mcp
Name	kcdc-mcp
Description	Demo project for Spring Boot
Package name	dev.danvega.kcdc
Packaging	● Jar ○ War
Java	● 24 ○ 21 ○ 17

Dependencies ADD DEPENDENCIES... ⌘ + B

Spring Web WEB

Build web, including RESTful, applications using Spring MVC. Uses Apache Tomcat as the default embedded container.

Model Context Protocol Server AI

Spring AI support for Model Context Protocol (MCP) servers.

GENERATE ⌘ + ↵ EXPLORE CTRL + SPACE ...

BUILDING AN MCP SERVER IN JAVA / SPRING

Spring Programming Model

With some additional MCP APIs

- Tools
- Prompts
- Resources

```java
@Component
public class SessionTools {

    private static final Logger log = LoggerFactory.getLogger(SessionTools.class);
    private Conference conference;
    private final ObjectMapper objectMapper;

    public SessionTools(ObjectMapper objectMapper) {
        this.objectMapper = objectMapper;
    }

    @McpTool(name = "codemash-get-conference-data", description = "Get all conference data including sessions, tracks, rooms and conference details")
    public Conference getConferenceData() {
        return conference;
    }

    @McpTool(name = "codemash-get-sessions", description = "Get a list of all sessions")
    public List<Session> getSessions() {
        return conference.sessions();
    }
```

```java
@SpringBootTest
class SessionToolsTest {

    @Autowired
    private SessionTools sessionTools;

    @Test
    void getConferenceDataReturnsValidData() {
        Conference conference = sessionTools.getConferenceData();

        assertThat(conference).isNotNull();
        assertThat(conference.name()).isEqualTo("Commit Your Code Conference (CYC25)");
        assertThat(conference.year()).isEqualTo(2025);
        assertThat(conference.dates()).containsExactly("2025-09-25", "2025-09-26");
        assertThat(conference.location()).isEqualTo("Yum! Brands International, Plano, TX");

        assertThat(conference.tracks()).isNotEmpty();
        assertThat(conference.tracks()).contains("JavaScript", "Java", "Cloud", ".NET", "Leadership", "AI");

        assertThat(conference.rooms()).isNotEmpty();
        assertThat(conference.rooms()).contains("Red Room", "Yellow Room");

        assertThat(conference.sessions()).isNotEmpty();

        Session firstSession = conference.sessions().get(0);
        assertThat(firstSession).isNotNull();
        assertThat(firstSession.title()).isNotBlank();
        assertThat(firstSession.day()).isNotBlank();
        assertThat(firstSession.time()).isNotBlank();
        assertThat(firstSession.speakers()).isNotNull();

        boolean hasSessionWithSpeakers = conference.sessions().stream()
                .anyMatch(session -> session.speakers() != null && session.speakers().length > 0);
        assertThat(hasSessionWithSpeakers).isTrue();
    }
}
```

TESTING YOUR MCP SERVERS

- Create an executable JAR
- Test using an MCP Client
  - Spring MCP Client
  - Claude Desktop
  - Cursor / Windsurf / Junie
  - Any MCP Client of your choice

MCP Inspector v0.18.0

Transport Type

Streamable HTTP

URL

http://localhost:8081/mcp

Connection Type

Via Proxy

Server Entry | Servers File

Authentication

Configuration

Reconnect | Disconnect

Connected

codemash-mcp-server
Version: 1.0.0

Logging Level

debug

System

Resources | Prompts | Tools | Ping | Sampling | Elicitations | Roots | Auth | Metadata

Tools

List Tools

Clear

codemash-sessions-by-date
Returns the count of sessions by date

codemash-sessions-by-track
Returns the count of sessions for a specific track

codemash-get-conference-data
Get all conference data including sessions, tracks, rooms and conference details

codemash-get-sessions
Get a list of all sessions

codemash-get-sessions

Get a list of all sessions

Tool-specific Metadata:

No metadata pairs.

Run Tool

Copy Input

Tool Result: Success


```
[
  0: {
    day: "2026-01-13"
    time: "07:00 AM"
    duration: "1h"
    title: "Breakfast (Tuesday)"
    type: "Meal"
    speakers: []
    room: "Kilimanjaro"
    track: [
      0: "Meals & Entertainment"
    ]
  }
  1: {
    day: "2026-01-13"
```


History

Clear

4. tools/call
3. tools/list
2. logging/setLevel
1. initialize

Server Notifications

No notifications yet