

The AI Alphabet Soup!



Isabelle MAUNY
EU Field CTO - WSO2



An AI story...

First we had **LLMs**. We wrote stories thanks to **GPTs**, but soon needed **RAG** and **CAG** to create more relevant content.

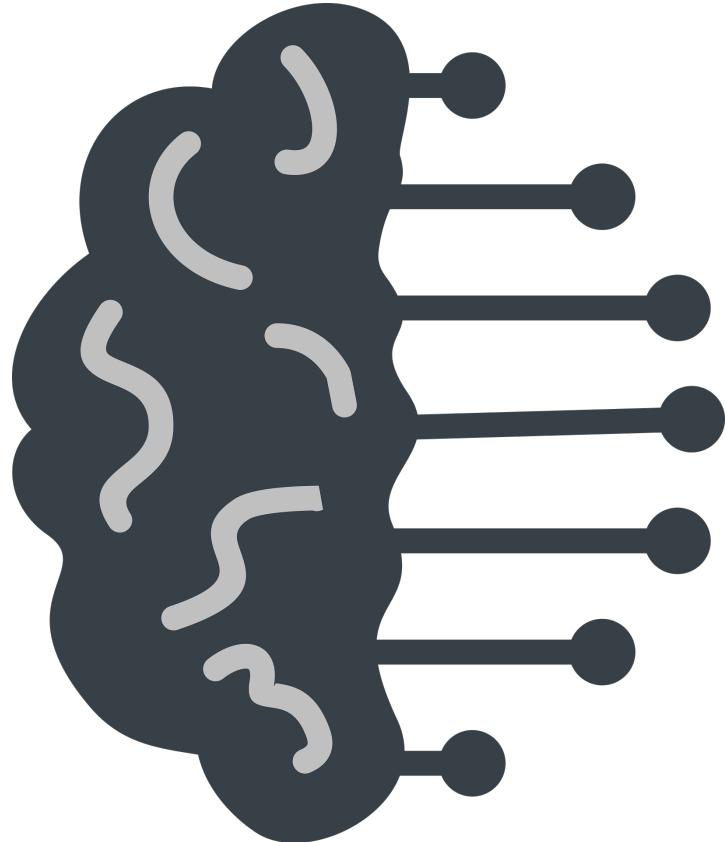
We then realized **APIs** were not enough and created **MCPs**, so that agents can use **tools** to execute autonomously.

To control our LLM consumption, we had to install an **AI gateway** which allowed us to set **guardrails** and manage our **tokens** usage!

Somewhere along the way, we almost forgot about security but **OAuth2** came to the rescue!

Core building blocks: LLMs / GPTs...

- LLMs (Large Language Models) are trained on a very large amount of data.
- GPTs (Generative Pre-trained Transformer) are a special type of LLMs
 - **Generative**: create new text
 - **Pre-trained**
 - **Transformer**: using a transformer architecture.
At a high level: GPTs are good at predicting what the next **token** in a response will be based on **context**.
- **GenAI** (Generative AI) : AI technology focused on generating content (text, images, audio, video, etc.)



...and tokens!

OpenAI Platform

Docs API reference Log in Sign up

model, and the total count of tokens in that piece of text.

GPT-4o & GPT-4o mini GPT-3.5 & GPT-4 GPT-3 (Legacy)

OpenAI's large language models process text using tokens, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens

Clear Show example

Tokens	Characters
48	267

OpenAI's large language models process text using tokens, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens

Let's walk an example...

Planning a trip to Japan...





An application that uses Gen AI models to generate content like text and images

IM I want to spend a few days in Japan

Japan is an incredible destination! A few days will give you a great taste of the country, though you'll probably want to return for more.

Popular options for short trips:

Tokyo focus - Perfect for first-time visitors. You could explore different neighborhoods like Shibuya, Harajuku, Asakusa, and Ginza. Visit temples like Senso-ji, experience the fish market area, and enjoy the food scene from street food to high-end dining.

Tokyo + Kyoto - If you can manage it, the shinkansen (bullet train) makes this combination very doable. Tokyo for modern Japan, Kyoto for traditional temples, gardens, and historic districts like Gion.

Osaka base - Great for food lovers, with easy day trips to Kyoto or Nara to see the deer park.

Practical considerations:

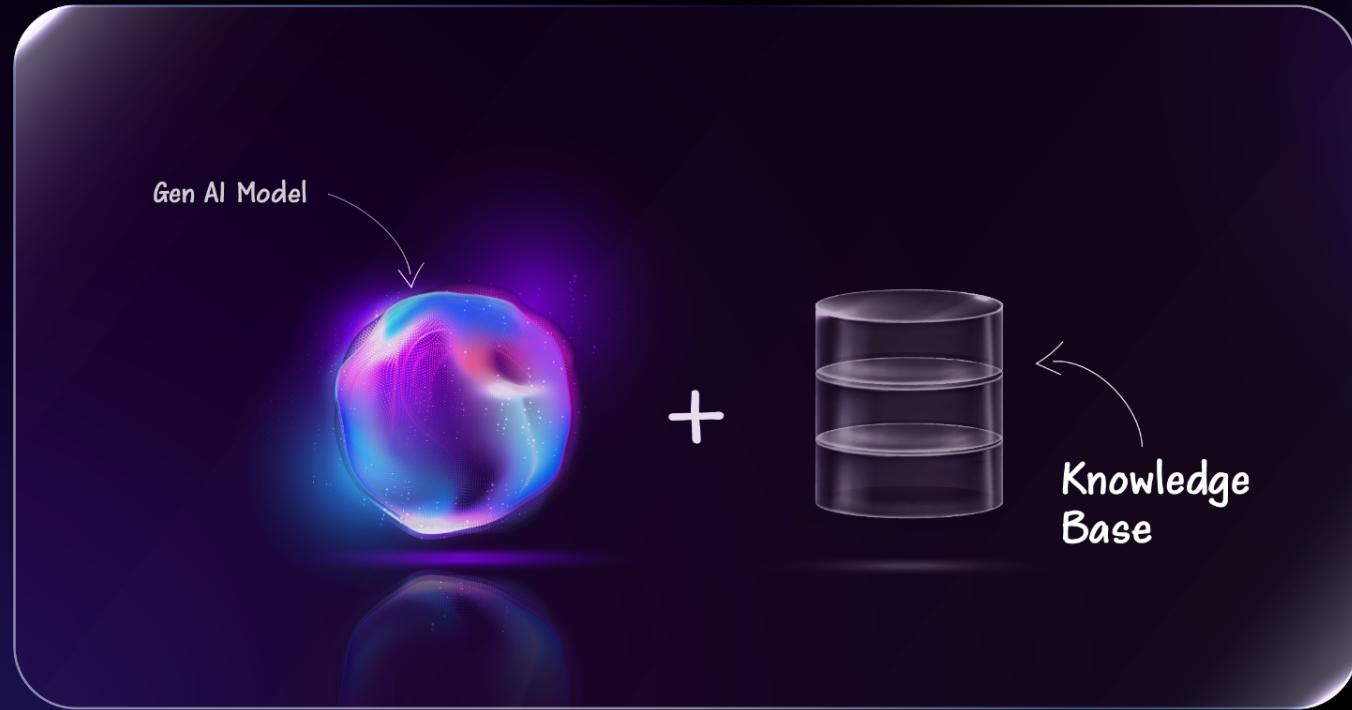
- JR Pass might be worth it if you're traveling between cities
- Many signs have English, especially in major tourist areas

GENERIC RESPONSE BASED ON TRAINING DATA

Add more context to the prompt !

What kind of experience are you most drawn to? The bustling modern city life, traditional culture and temples, food adventures, or a mix of everything? Also, what time of year are you thinking of visiting? This can really influence what you'll want to prioritize.

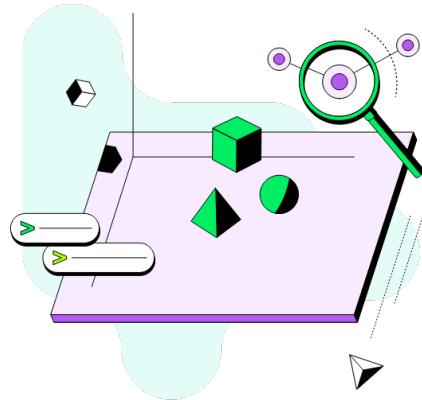
- This is a **User Experience issue**: if you're a *vacation.com*, you can't ask this from every customer, especially since you already have a lot of context
- You could use **CAG (Cache Augmented Generation)**: you could pass full documents as context, **but**
 - Context of LLMs is currently limited by number of input tokens (50k-100k)
 - Beware of \$\$\$ / €€€ !
- Right choice in some cases though!



An application that uses Retrieval Augmented Generation (RAG)

Adding more context through RAG

- Can use many sources of data
 - Documentation
 - Databases
 - Customer information
 - Emails
- Data must be careful prepared, indexed, chunked => **Vector Databases**



<https://www.mongodb.com/products/platform/atlas-vector-search>



Based on Alex's travel profile, I'd recommend a **Kyoto-focused cultural immersion trip** since they already experienced Tokyo thoroughly and rated it 10/10 for culture and food.

Recommended Itinerary: Traditional Japan (7-8 days)

Kyoto (5 days) + Nara (1 day) + Hakone (2 days)

Why This Fits Alex's Profile:

Accommodation: Stay in traditional ryokans - this matches their love for heritage hotels (India) and riads (Morocco). Budget-wise, quality ryokans run \$200-350/night, fitting their range.

Cultural Immersion:

- Kyoto's temples and traditional districts (Gion, Arashiyama bamboo grove)
- Nara's less crowded, authentic temple experience
- Traditional craft workshops (pottery, textile dyeing) - similar to their loved block printing class in India

CUSTOMIZED RESPONSE BASED ON ALEX'S PROFILE

Ok! Let's Go !



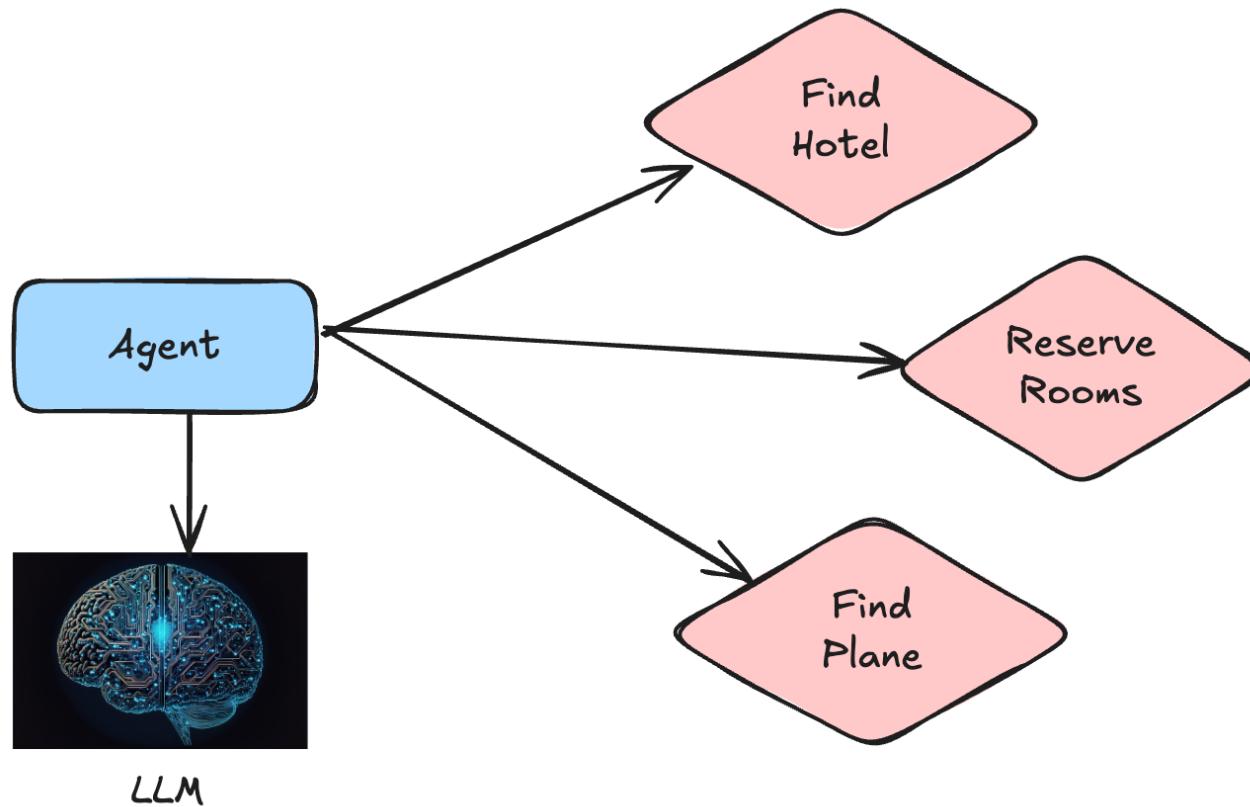
So far, we are in GenAI mode, not in Agent mode.

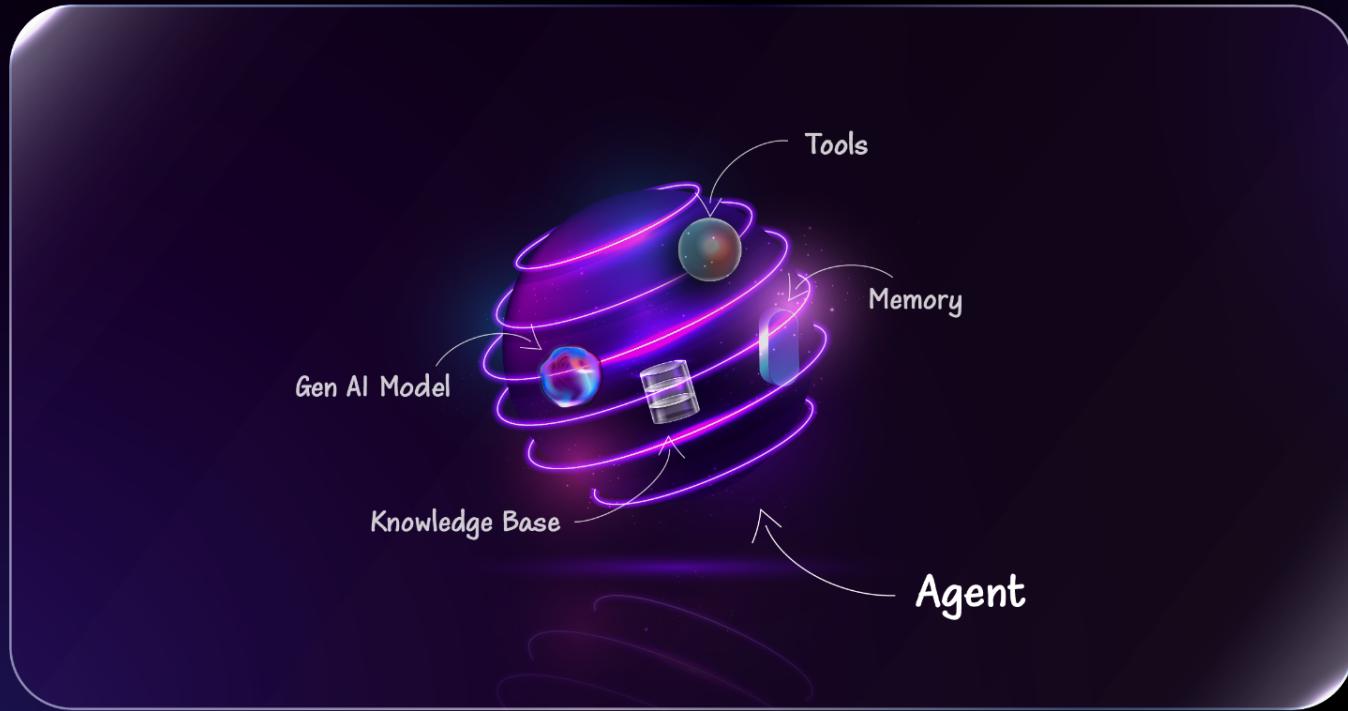
IM

Can you make reservations for me?

I can't actually make reservations for you - I don't have access to booking systems or the ability to make real-world transactions on your behalf.

We need tools!





An application that uses an Agent



Tools are just functions!

Diagram / Diagram

← AI Chat Agent POST chat (request: ai:ChatReqMessage) ↶ ai:ChatRespMessage|error

The screenshot shows a workflow editor for an AI Chat Agent. On the left, the main workspace displays a flowchart starting with a 'Start' node, followed by a function call node 'f log : printInfo "Starting agent"', then an 'AI Agent llmResponse' node. This node contains a 'Memory ai:MessageWindowChatMemory' component and a descriptive text block: 'Music recommendations assistant. You are a friendly companion recommending playlists depending on the weather forecast in the city the user lives in. You can fetch weather data and suggest appropriate playlists based on the conditions.' Below this text is a 'Toolbox' section with three items: 'weatherforecast', 'getPlaylistSuggestions', and a plus sign icon. A red box highlights the 'Add Tool' button at the top right of the toolbox. On the far left, there are navigation icons: a house, a plus sign, and a minus sign.

Add Tool

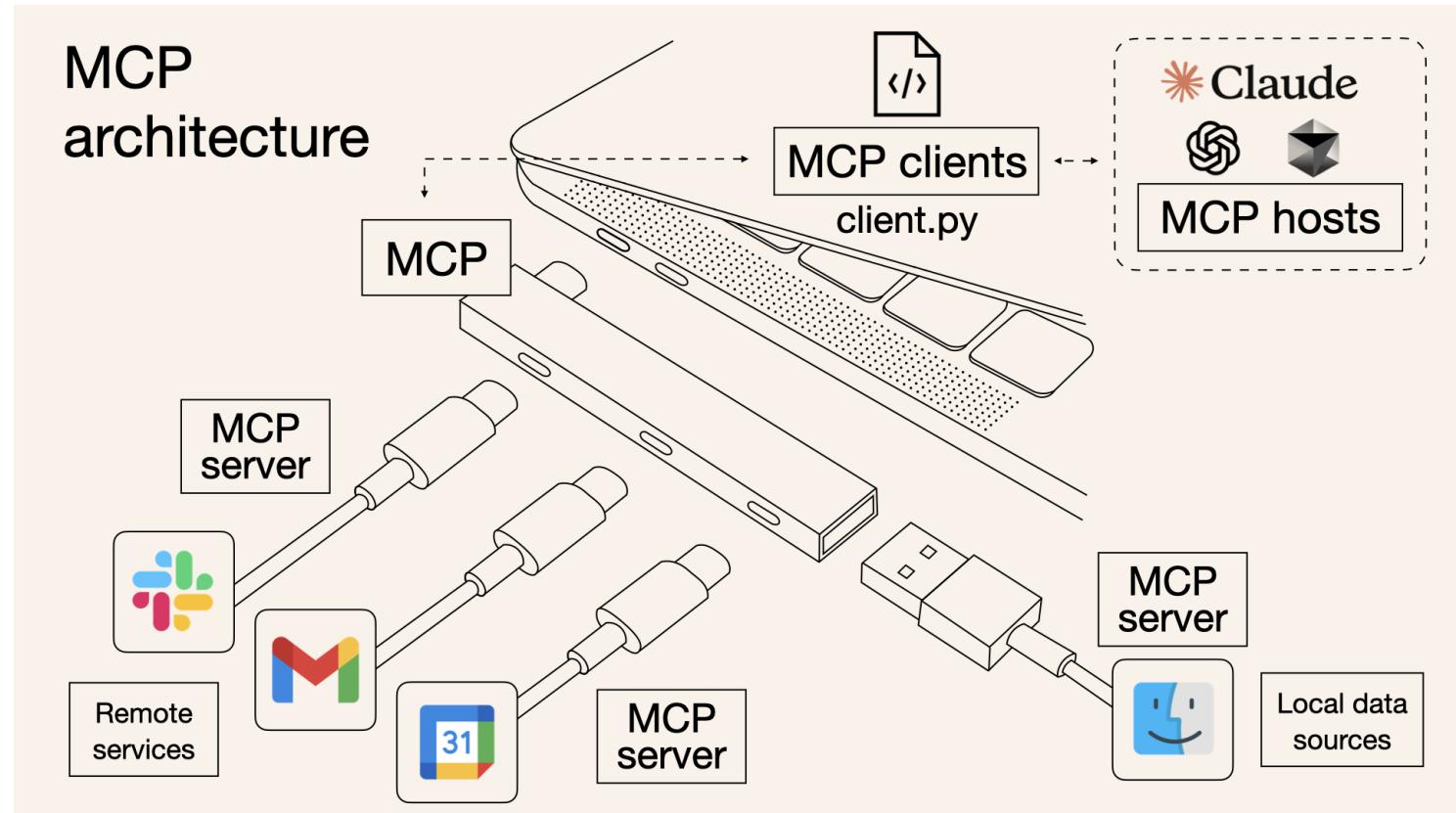
Choose a tool to add to the Agent or create a new one.

Tools

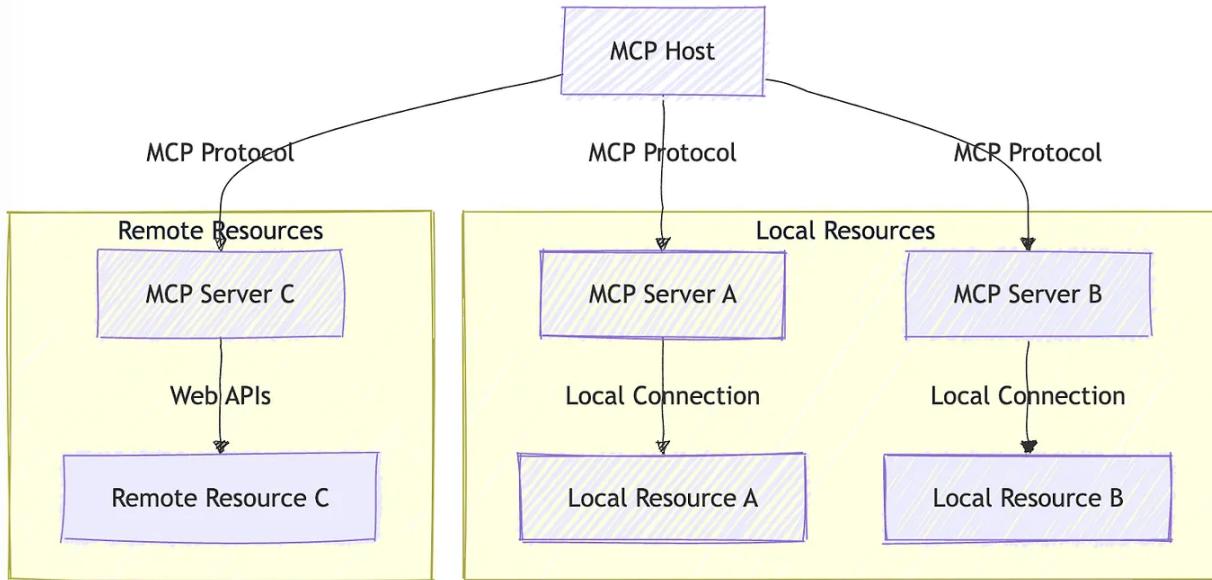
- weatherforecast
- getPlaylistSuggestions

+

How do we get an LLM to talk to any tool, in a uniform way ?



MCP Architecture



The MCP Trivia!

- Do we need MCPs to have tools ? **No.**
- Do we need APIs to write MCPs ? **No.**
- Do we need tools to build agents ? **Yes.**
- Do we need to secure MCPs ? **Yes. Yes and Yes.**





Big tendency now: offer management API as MCP

□ README 🤝 Code of conduct ⚙️ MIT license ⚙️ Security [Edit](#) [☰](#)

GitHub MCP Server

The GitHub MCP Server is a [Model Context Protocol \(MCP\)](#) server that provides seamless integration with GitHub APIs, enabling advanced automation and interaction capabilities for developers and tools.

Use Cases

- Automating GitHub workflows and processes.
- Extracting and analyzing data from GitHub repositories.
- Building AI powered tools and applications that interact with GitHub's ecosystem.

Remote GitHub MCP Server

[VS Code Install Server](#) [VS Code Insiders Install Server](#)

BUUUUUT !

2025-05-26

GitHub MCP Exploited: Accessing private repositories via MCP

We showcase a critical vulnerability with the official GitHub MCP server, allowing attackers to access private repository data. The vulnerability is among the first discovered by Invariant's security analyzer for detecting toxic agent flows.



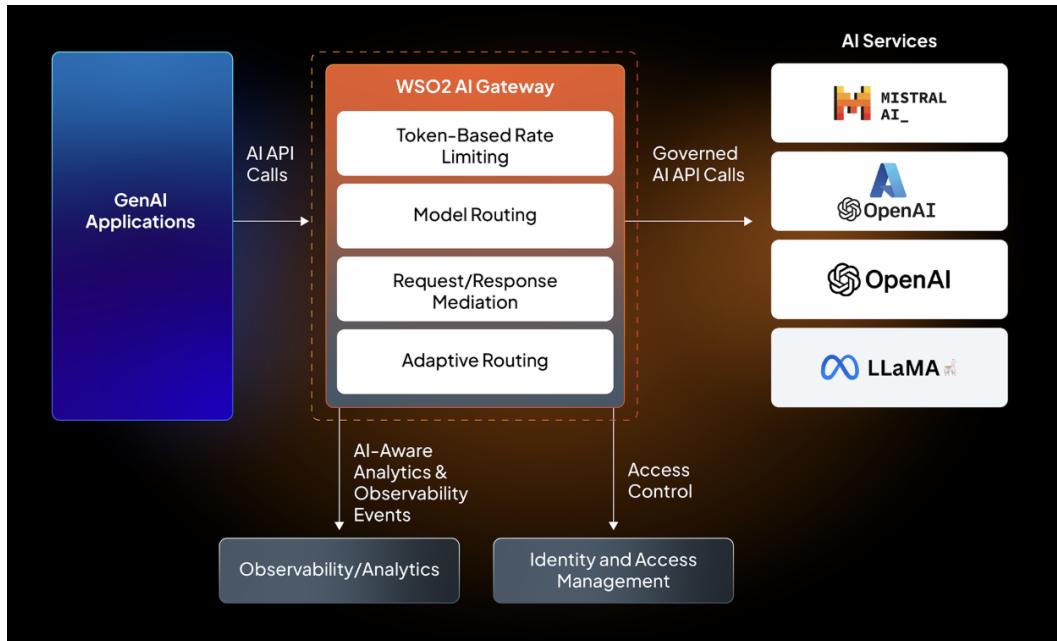
MCP and Security !

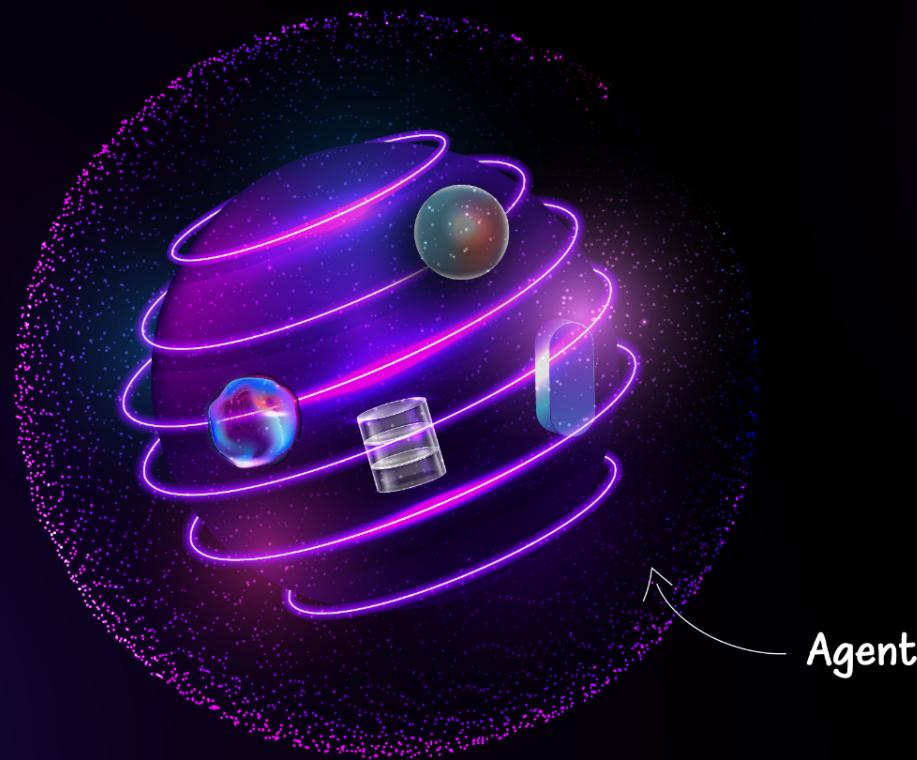
- Adding an MCP on top of an insecure API is a recipe for disaster.
- MCP require authentication and authorization
- Inputs must be validated
- Caution: Harder to test today than an API



Why an AI Gateway ?

- **AI guardrails and prompt engineering with request/response mediation**
- **Token-based rate limiting and analytics**
- **Multi-provider model routing**
- **Adaptive request handling**





Agent

An application where the Agent
is the application

The next level ! Autonomous agents

[Legend](#)[Processus](#)[F.A.Q](#)[Agents IA](#)[Charly](#)[Intégrations](#)[Nos Offres](#)[Se connecter](#)[Reserver une démo](#)

Agent Téléphonie

Tom

Offrez à vos clients une expérience téléphonique ultra réactive, fluide et 100 % automatisée. Tom, notre agent IA...

[Voir les détails ↗](#)

Agent Marketing

John

Votre assistant IA pour des réseaux sociaux qui tournent tout seuls. John crée vos visuels, rédige vos posts et publie à votre...

[Voir les détails ↗](#)

Agent Relation Client

Mickael

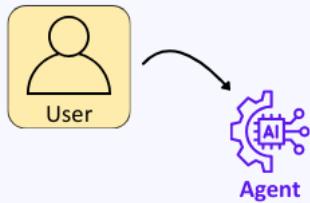
L'agent IA qui répond à vos clients, en continu, sur WhatsApp et sur votre site. Mickael est votre chatbot relation client...

[Voir les détails ↗](#)

Agents should have an identity ?

Interactive agent

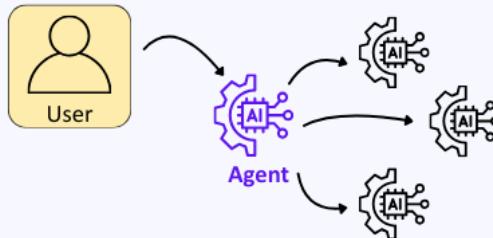
Single, specific,
repetitive, defined task



Tool calling for task

Autonomous agent

Complex, goal oriented, creates
plans to achieve outcome



Agent to agent collaboration

Digital employee

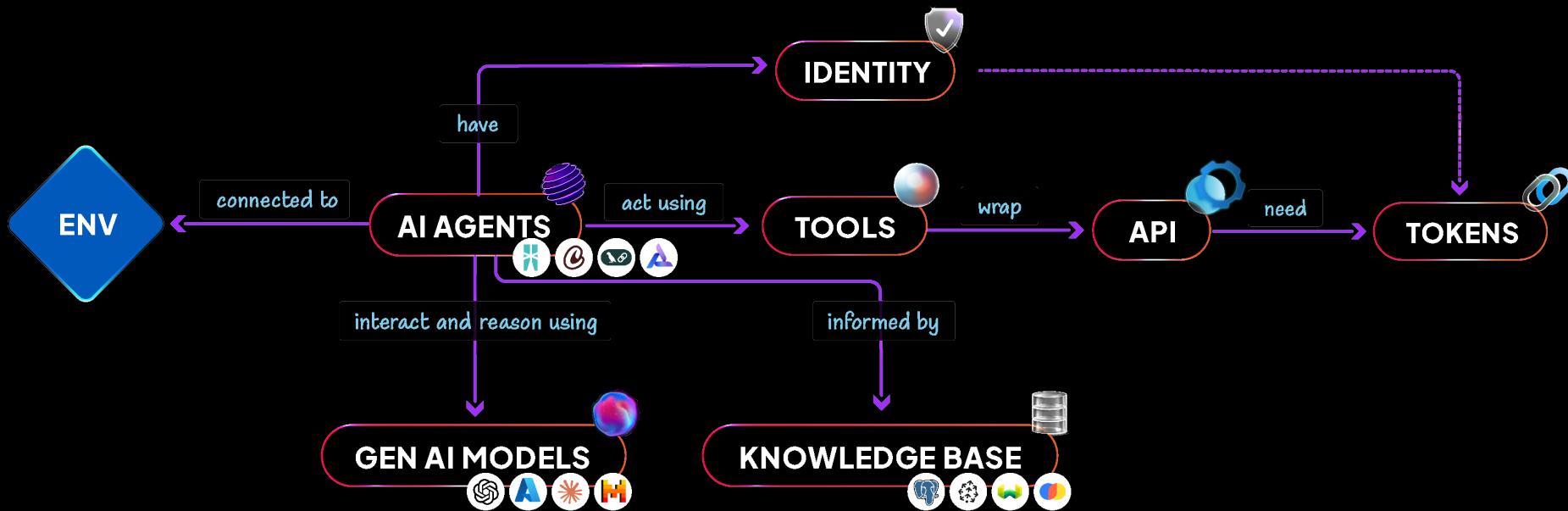
Learning driven, emulate
human decision making



Agent has an email address,
can chat with you in Teams,
has its own OneDrive, can
sign into a PC, etc.

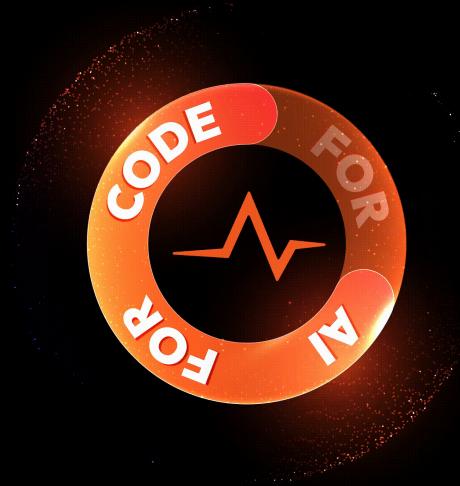
Agent in HR system, with role-based
access like a human on that team

PUTTING IT ALL TOGETHER!



AI is fundamentally changing software development

Developers productivity is supercharged through **AI for code** across the software development lifecycle.



Building AI-native applications is changing **code for AI** through emerging programming abstractions and building blocks

Blending language and code with
natural programming

WSO2CONASIA

July 29–31, 2025 | Colombo, Sri Lanka

Platformless Modernization and Celebrating 20 Years of Global Impact

- Inspiring keynotes
- Focused agenda
- Interactive sessions
- Customer stories
- Look to the future
- Networking

Full Conference

LKR 50,000 (150 €)

Includes 3-day conference access (July 29–31), meals during conference hours, networking sessions, and entry to all social events.

Full Conference + 4 Nights' Accommodation

LKR 200,000 (600 €)

Includes everything in the Full Conference package, plus 4 nights' accommodation at the conference venue (July 28–August 1).

[Register Now](#)



Gracias



wso2.com



Learning more

- **Vectors, Tokens and Embeddings:** <https://thenewstack.io/the-building-blocks-of-langs-vectors-tokens-and-embeddings/>
- **RAG vs CAG** from IBM Technology : <https://www.youtube.com/watch?v=HdafI0t3sEY>
- **MCP Tutorial by Anthropic:** <https://modelcontextprotocol.io/quickstart/server>
- **Transformers Architecture Insights:** <https://www.youtube.com/watch?v=wjZofJX0v4M>