The State of GenAI & ML in the Cloud Native Ecosystem

Alejandro Saucedo
Bartosz Ocytko

zalando

zalando

About us

Alejandro Saucedo

Director of Eng, Science & Product

Bartosz Ocytko

Executive Principal Engineer

Zalando | Who We Are

Leading European fashion.

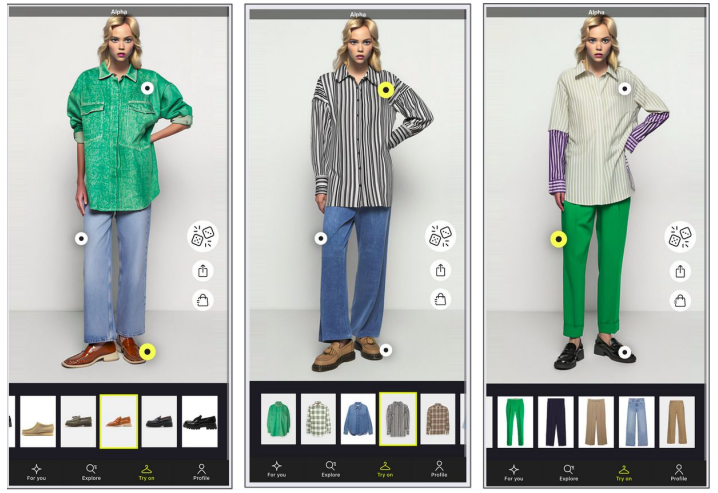
~15bn
Euro GMV

~51m
Active Customers

~2m
Articles

Zalando | GenAI & ML

Across ~80 Applied Science teams



Search & Recommendations	Warehouse Inventory Optimization	Conversational AI **GenAI**
Replenishment Prediction	Pricing Elasticities	Forecasting Demand, Returns, etc
Size & Fit Recommendations	Machine Translation **GenAI**	Fraud Detection
Virtual Clothing Try-On **GenAI**	Marketing Intelligence	Competitive Matching

We’re hiring!
Come work with us.

Find your career

Search for a role...

All open roles

jobs.zalando.com

The State of GenAI & ML in Cloud Native Ecosystem

Motivations

Market trends, opportunities and challenges

GenAI Products

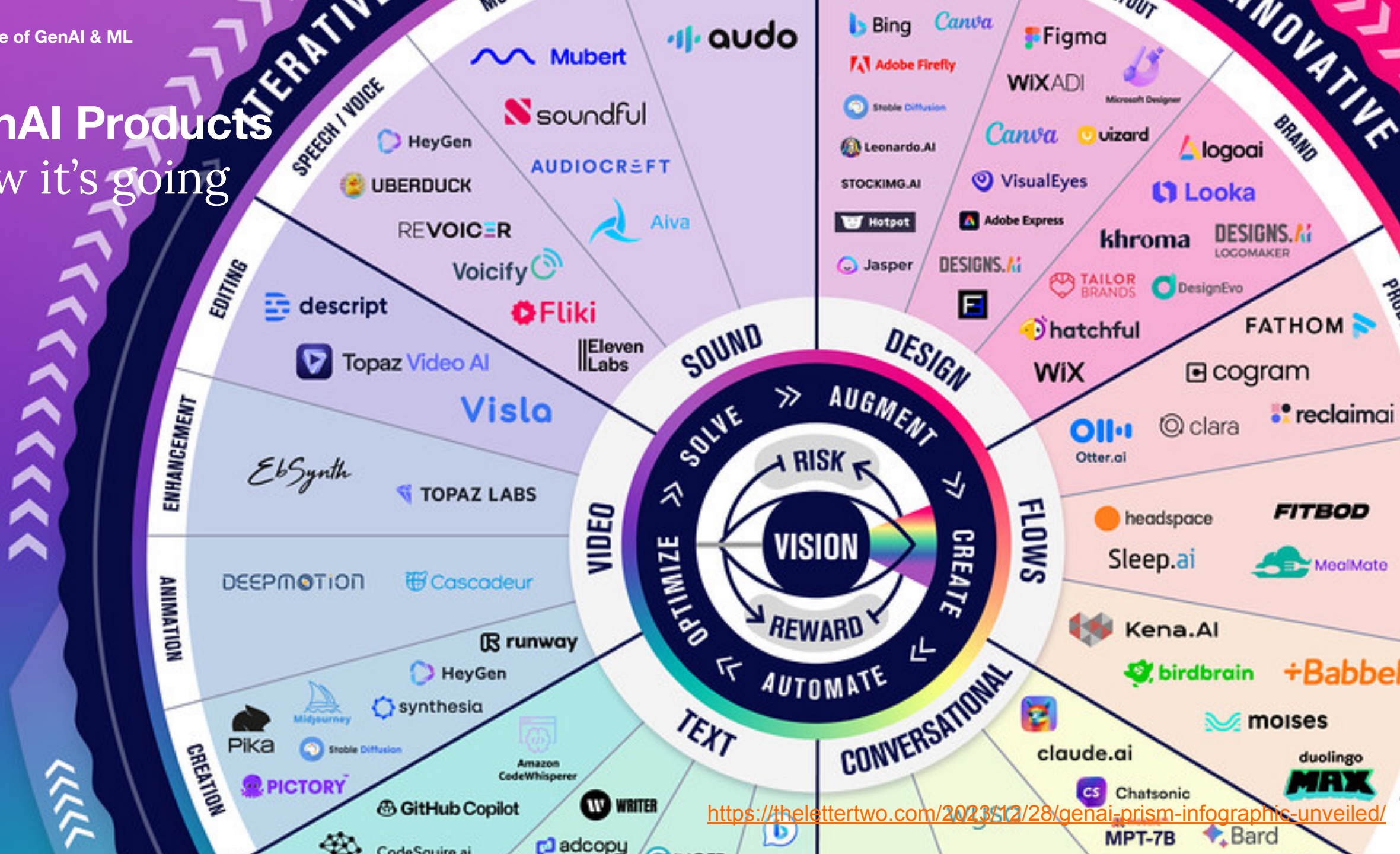
How it started

ChatGPT

DALL·E

Stable Diffusion

GenAI Products
How it's going



https://thelettertwo.com/2023/12/28/genai-prism-infographic-unveiled/

GenAI Products
How it's going

securityJustInterferesWithVibes

Meme

leo
@leojr94_

my saas was built with Cursor, zero hand written code

AI is no longer just an assistant, it's also the builder

Now, you can continue to whine about it or start building.

P.S. Yes, people pay for it

4:34 am · 15 Mar 2025 · **52.2K** Views

leo
@leojr94_

guys, i'm under attack

ever since I started to share how I built my SaaS using Cursor

random thing are happening, maxed out usage on api keys, people bypassing the subscription, creating random shit on db

as you know, I'm not technical so this is taking me longer that usual to figure out

for now, I will stop sharing what I do publicly on X

there are just some weird ppl out there

9:04 am · 17 Mar 2025 · **53.6K** Views

in 3 to 6 months, AI
ode software
of

Share Save

Shweta
@shweta_ai

I turned myself into an AI model! 🤖✨

Everything - the digital me, the outfits, the image-to-video transformation - is 100% AI-generated.

Here's exactly how I did it using Doji, Channel42, Kling AI & fal. 🧵👇

5:37 PM · Feb 27, 2025 · **194.2K** Views

Pres Mihaylov
@PreslavMihaylov

A non-engineer vibe-coded a simple book-suggestion app using Cursor & Claude, ended up extremely frustrated past the 85% mark.

"it feels like building an endless Jenga tower that just doesn't get higher"

The AI Market
and it's numbers.

~$390Bn.
Market valuation [1]

+5x.
Projected 5y increase [1]

+100m
Working in AI space by EOY [2]

[1] https://explodingtopics.com/blog/ai-statistics
[2] https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf

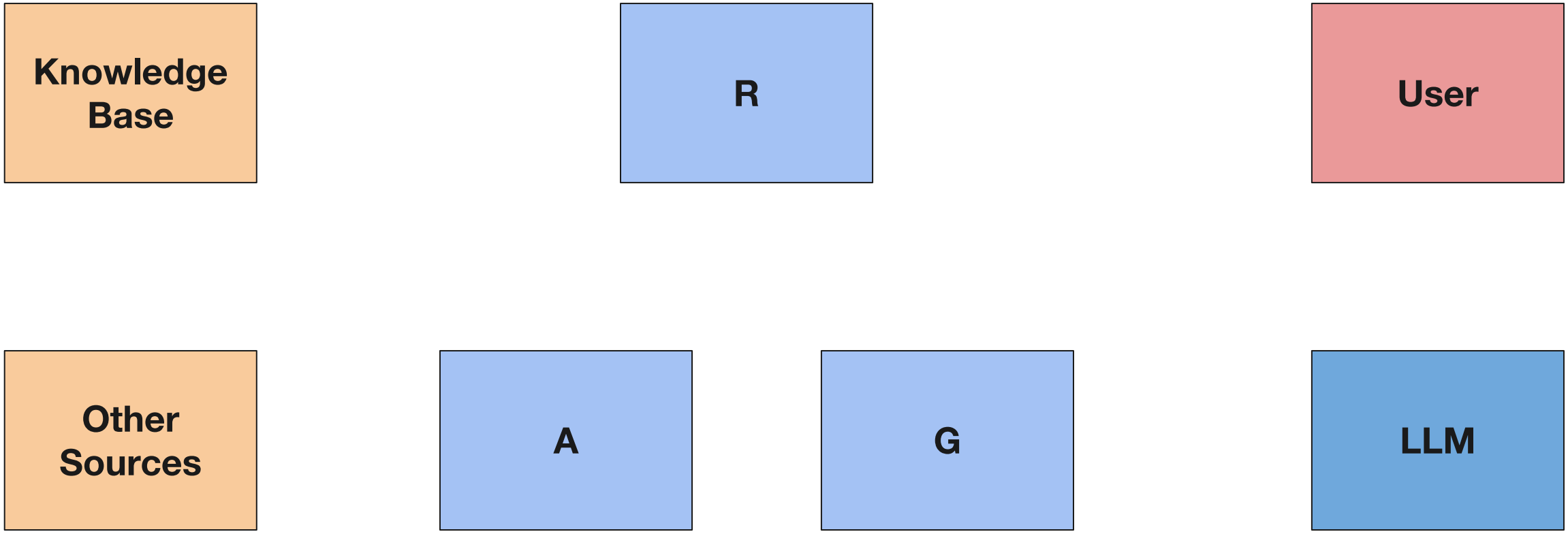
The State of GenAI & ML in Cloud Native Ecosystem

Challenges

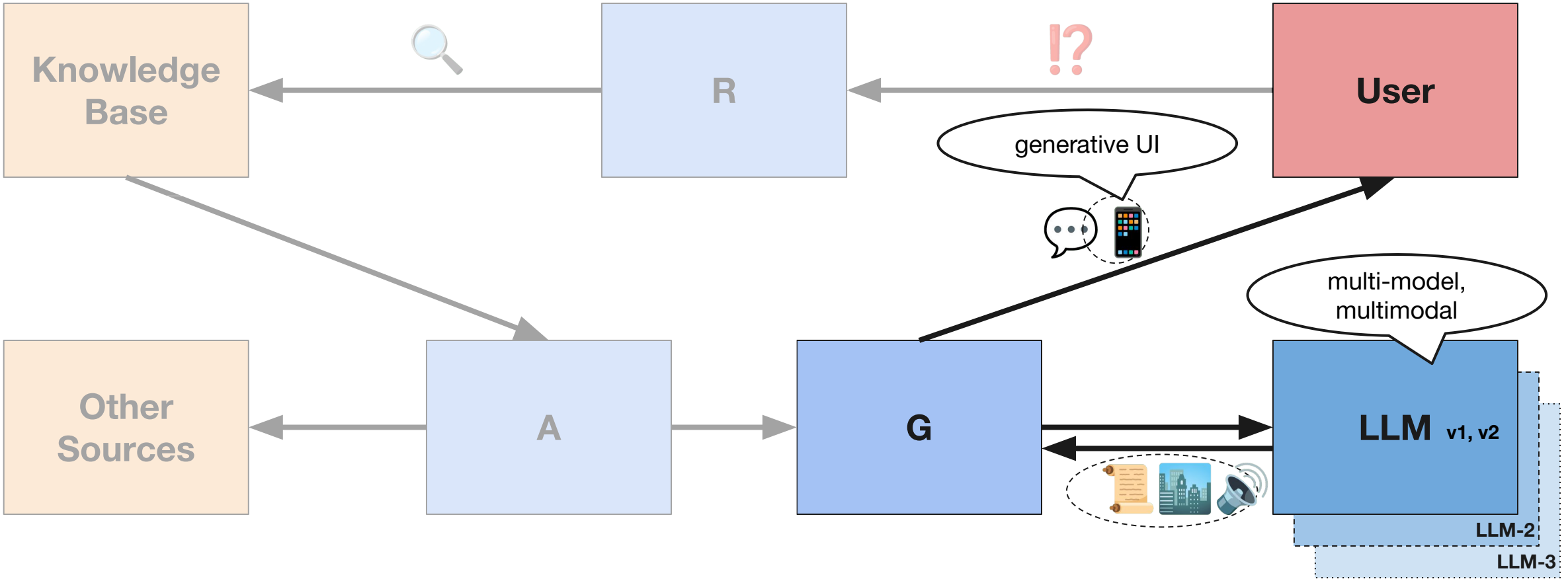
GenAI Journey Overview and Challenges reflected

RAG

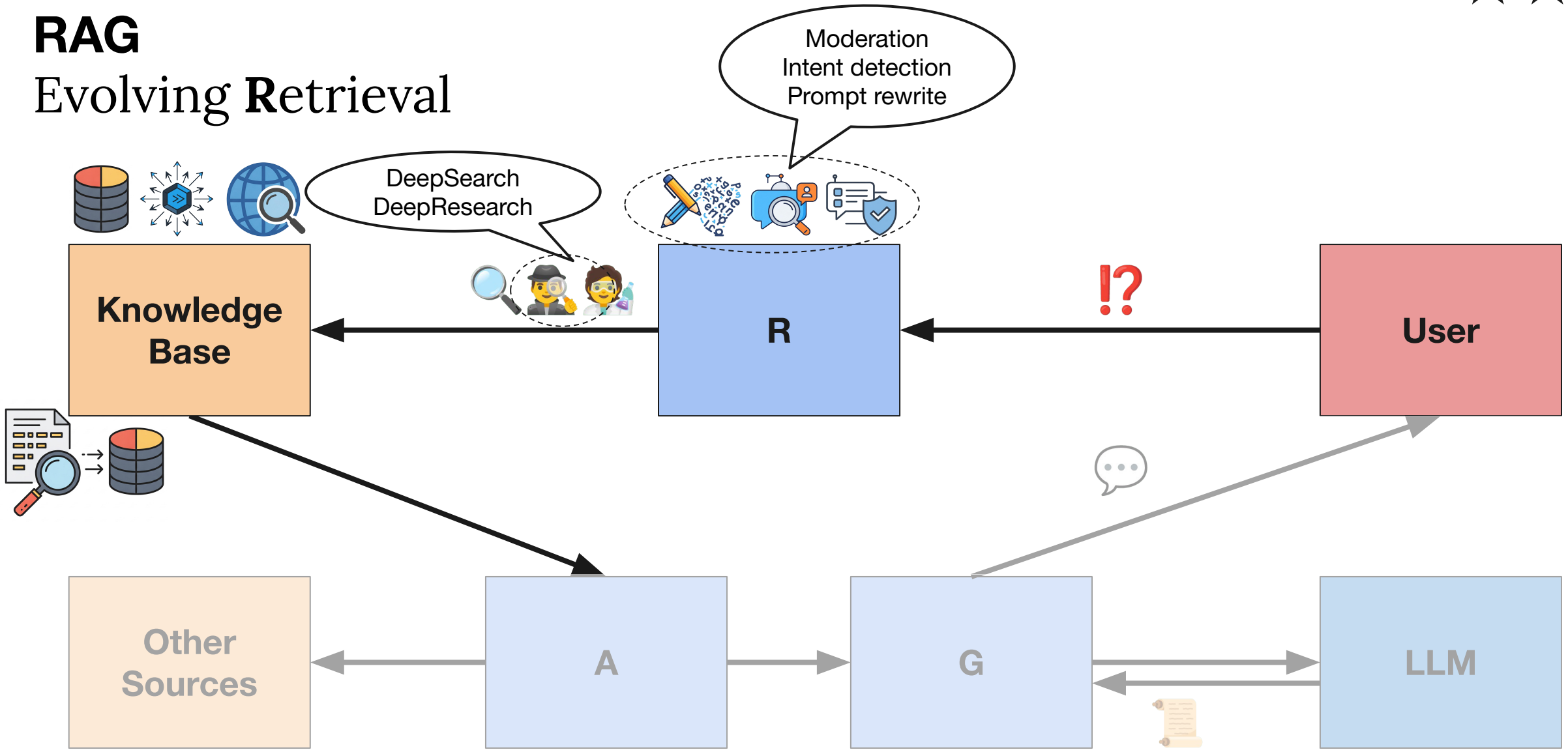
The “Hello World” for LLMs



RAG
Evolving Generation



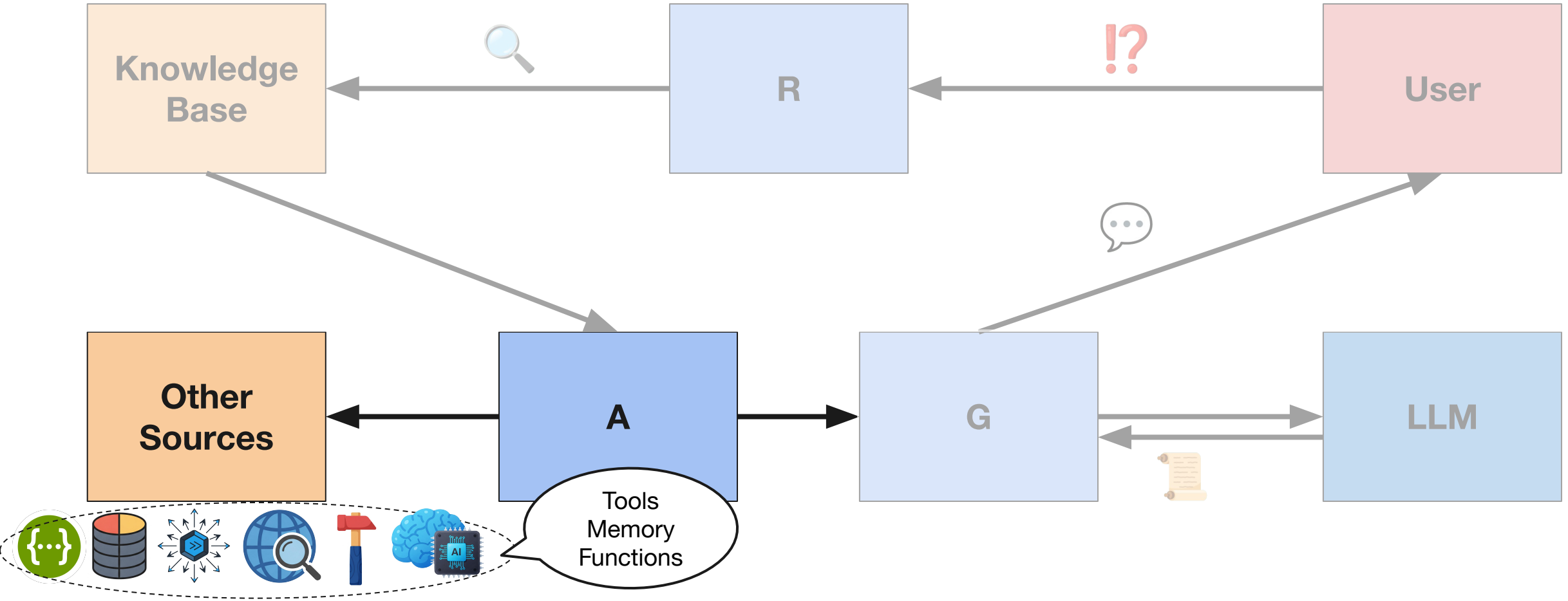
RAG
Evolving Retrieval



★★☆

RAG

Evolving Augmentation



RAG

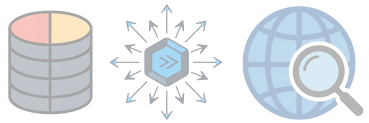
From RAG to Agents

Workflow orchestration

steps_limit,
recursion_limit

docker-compose
with N API keys

★★★

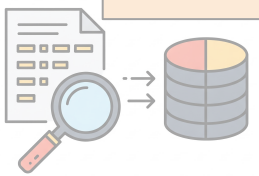


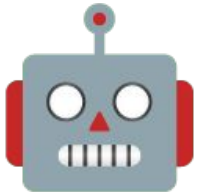
Knowledge Base

R

!?

User





Other Sources

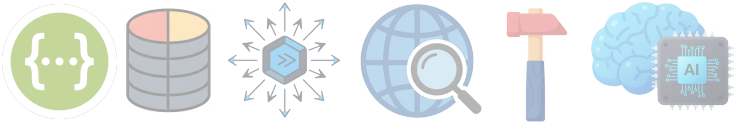
A

G

LLM

LLM-2

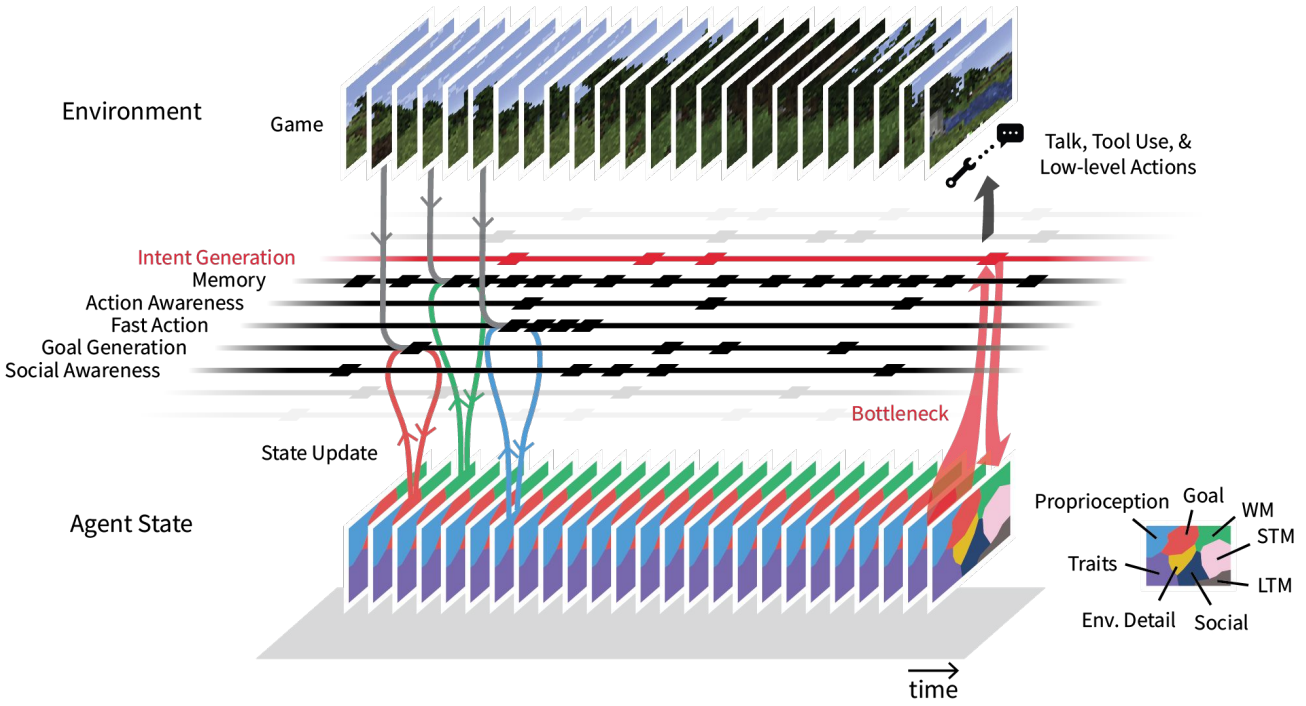
LLM-3



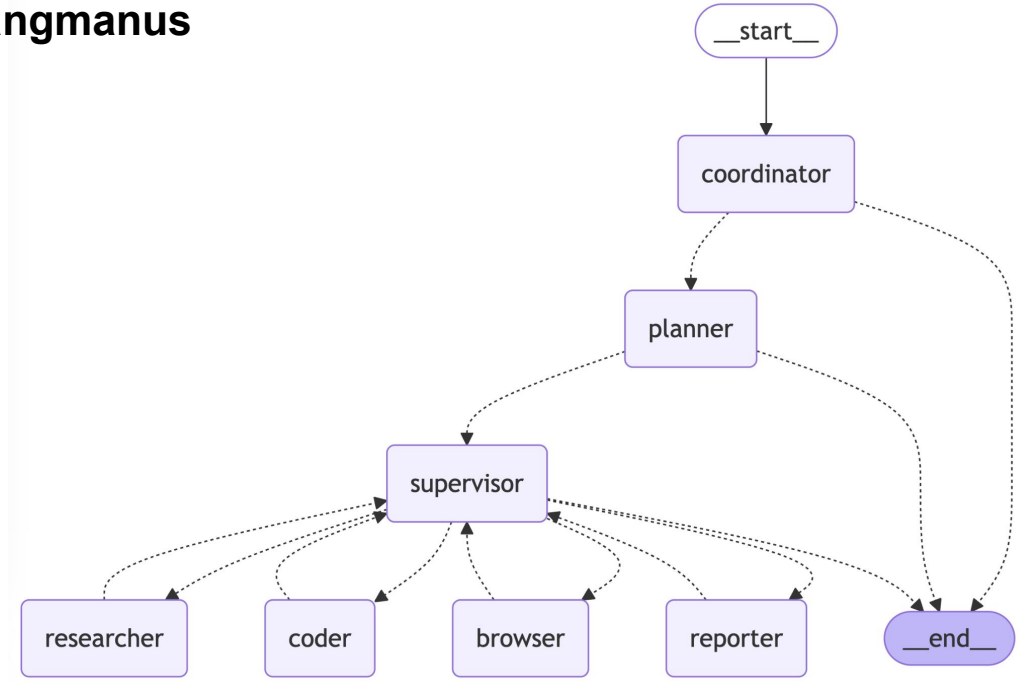
Agents on Steroids

Autonomous Agents

"Let loose in Minecraft, 1,000 autonomous AI agents collaborate to build their own society" (Project Sid from Altera)
https://digitalhumanity.substack.com/p/project-sid-many-agent-simulations

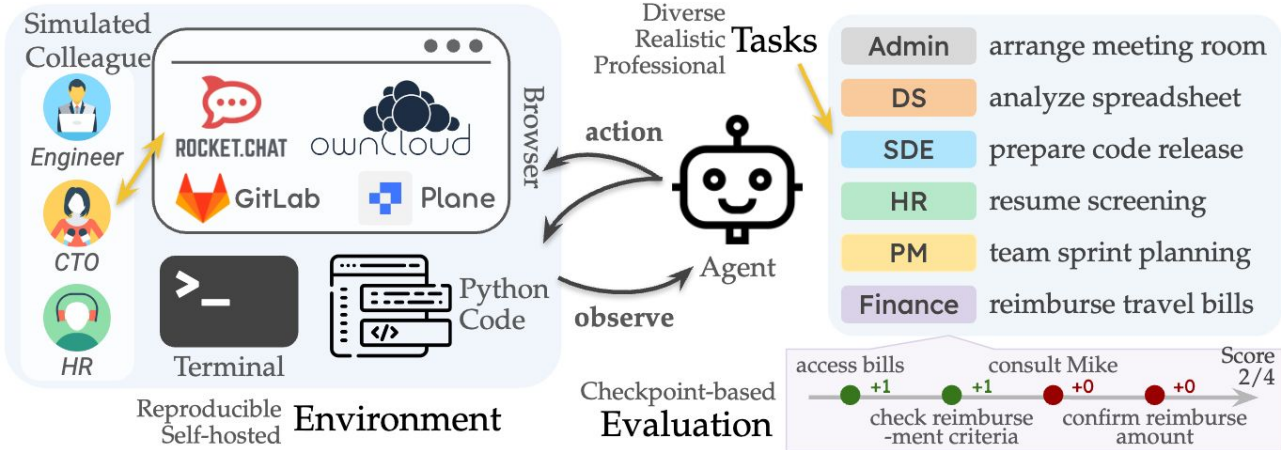


Langmanus



Benchmarks:

- **The Agent Company (Dec, 2024)**
- GAIA (Nov, 2023)
- AgentBench (v2, Oct, 2023)
- WebShop (v4, Feb 2023)



https://the-agent-company.com/

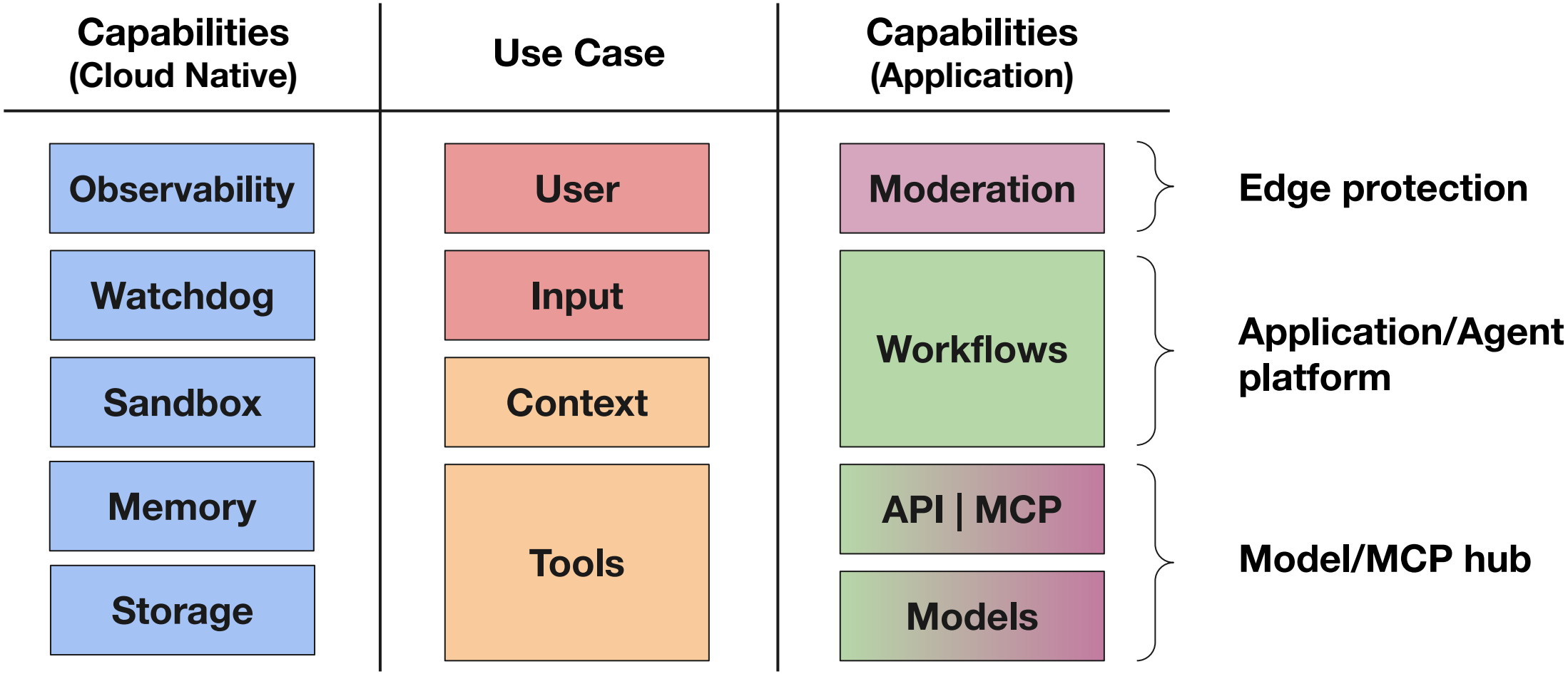
Challenges
per agency level

Agency level	Type	Challenges
☆☆☆	**Simple processor**	hallucinations, latency, model access, prompt quality, prompt injection, data poisoning
★☆☆	**Router**	limited capabilities: pre-defined, static workflows
★★☆	**Tool caller**	access to diverse tool libraries, **wrong tool**, params, invalid output
★★★	**Multi-step agents**	(non-) predictability, planning failures, **communication, large prompts**
★★★	**Multi-agent**	**security, rogue agents, endless loops**

agency definitions from https://huggingface.co/docs/smolagents/en/conceptual_guides/intro_agents

Cloud Native Agents

Capabilities



zalando

The State of GenAI & ML in Cloud Native Ecosystem

Best Practices

Overview of Best Practices in Industry

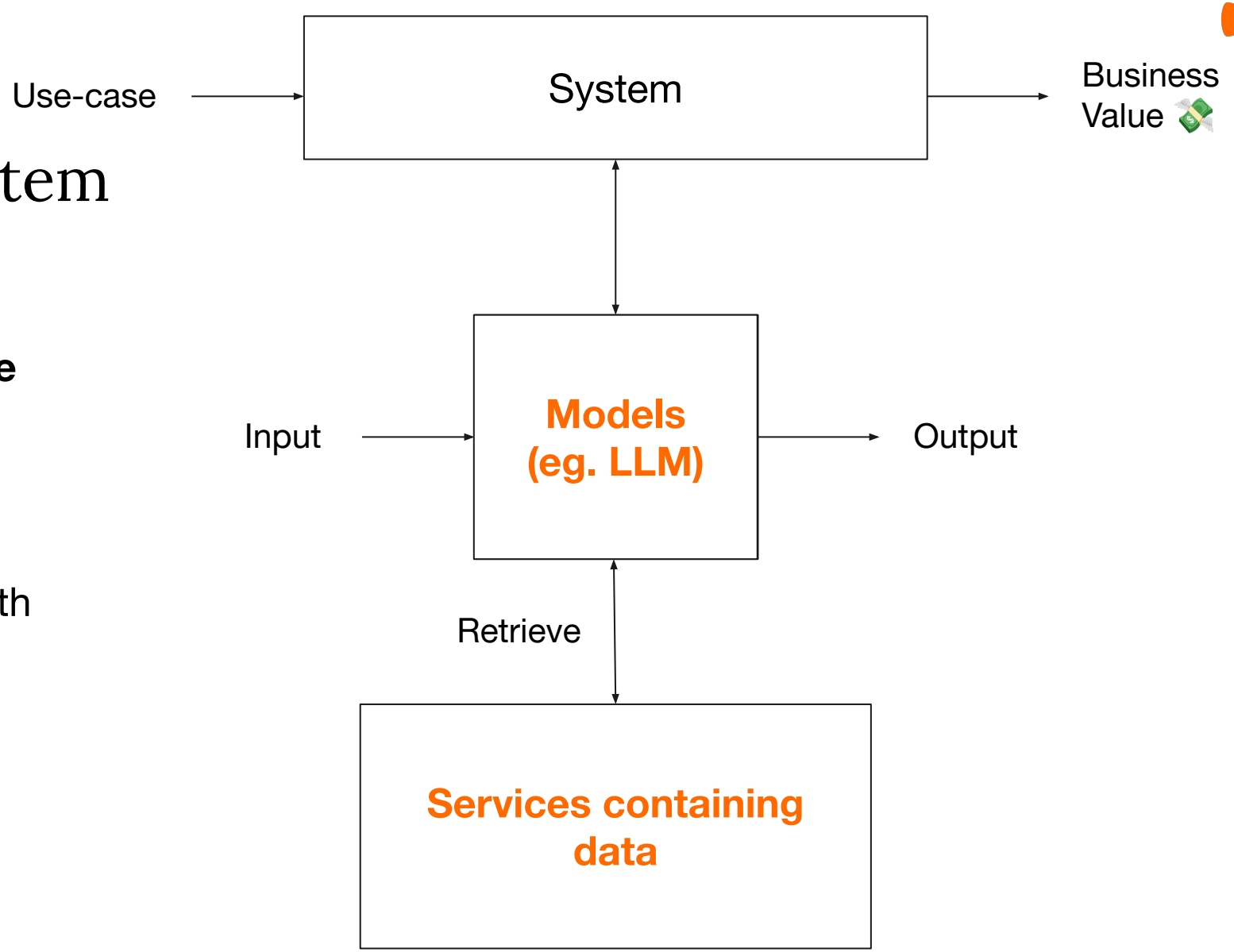
AI Ecosystem

From AI Model to AI System

Evolution from specific **applied science capabilities** to sophisticated **organisational capabilities.**

Encompassing **complex use-cases** with clear tangible **business value**.

Leveraging advancements in **software operations** applied into **AI delivery**.

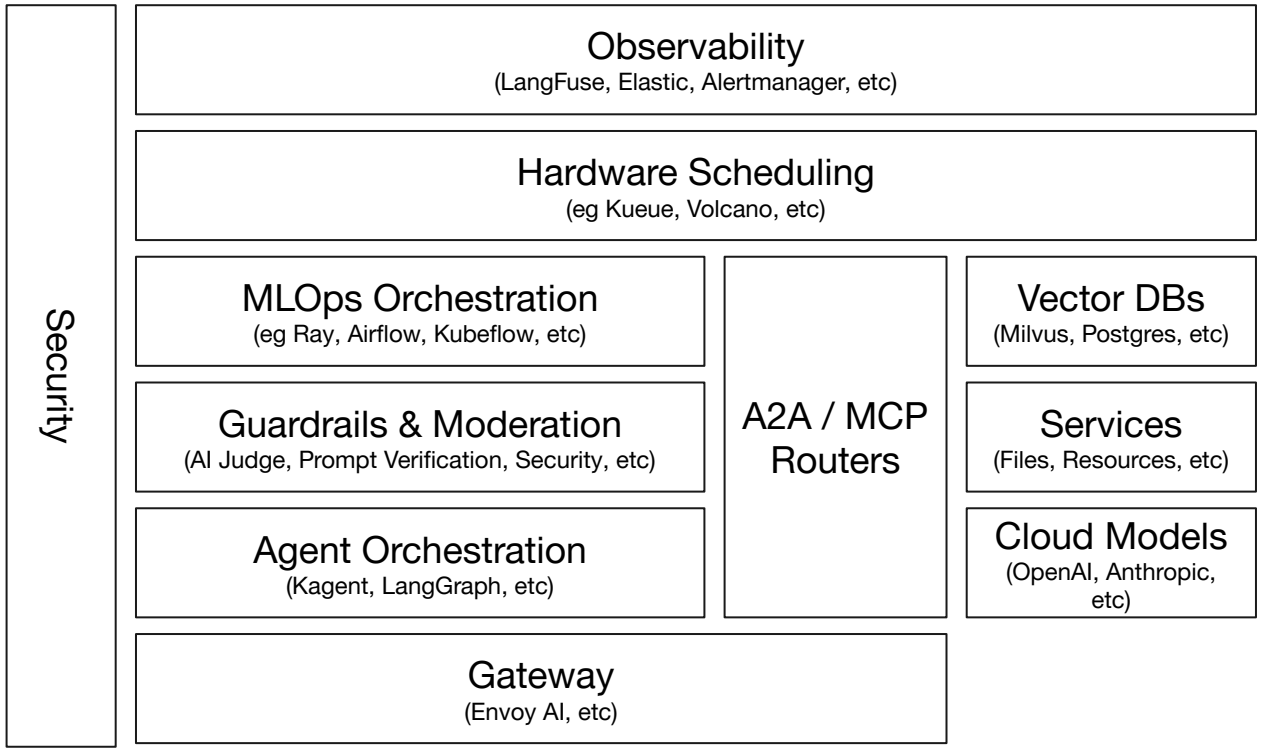


AI Ecosystem
From Model- to Data-Centric

The **core potential** is not on the models themselves, but in the **interconnected data-centric ecosystem**

This includes **complex orchestration** across **applications**, **infrastructure** and **complex data flow** in agentic systems

This extracts best practices from **traditional MLOps** but brings **further complexities** and **tooling requirements**

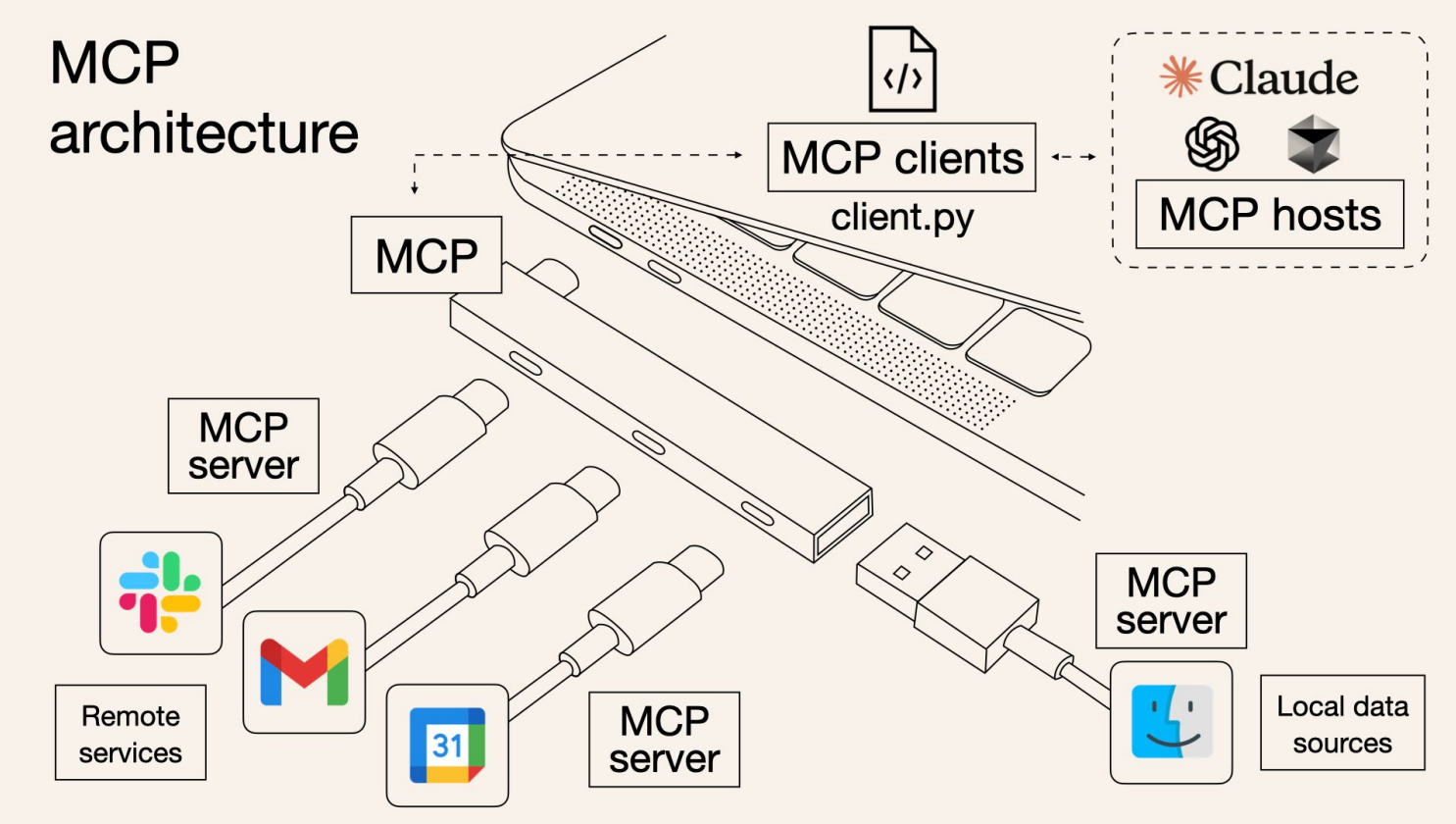


AI Ecosystem
The “Boring” Data

The archetypal **garbage-in = garbage-out** is true now more than ever across our intelligent systems and processes

This requires equally new protocols that drive **standardisation** and best practice for **interoperability** (eg envoy, MCP, etc)

Strategic investment in the “Boring” data is key to be **successful in the long term**, whilst **capturing value in the short term**.



ML Security
Considerations of Security

Taxonomy on Agents and Agentic Systems, both for **single- and multi-agent** contexts

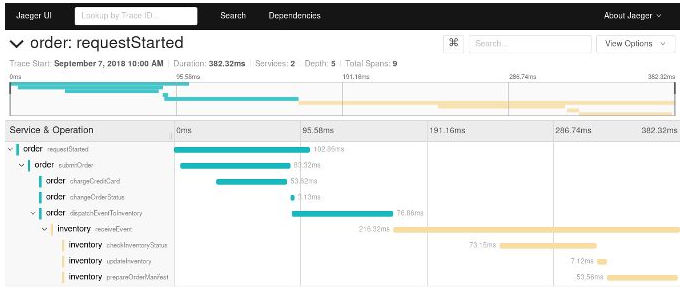
Taxonomy on Agentic System **Vulnerabilities** and **Exploits**

Examples include: Data poisoning (scrape data) Memory poisoning, remote code execution, identity spoofing, multi-agent chain exploits, cascading hallucination attacks, intent breaking, **etc**

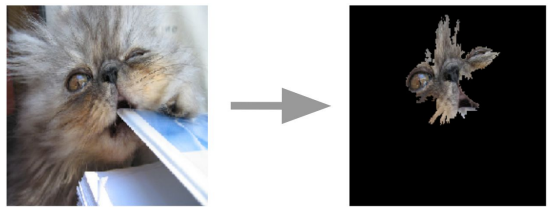
OWASP Agentic AI Threats & Mitigations: https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

MLOps Monitoring
Paradigms of Observability

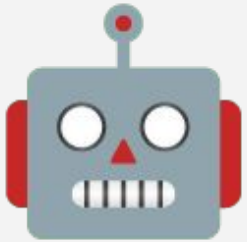
Service Performance



Explainability



Agentic System



Observability

LLM Perf / Debugging

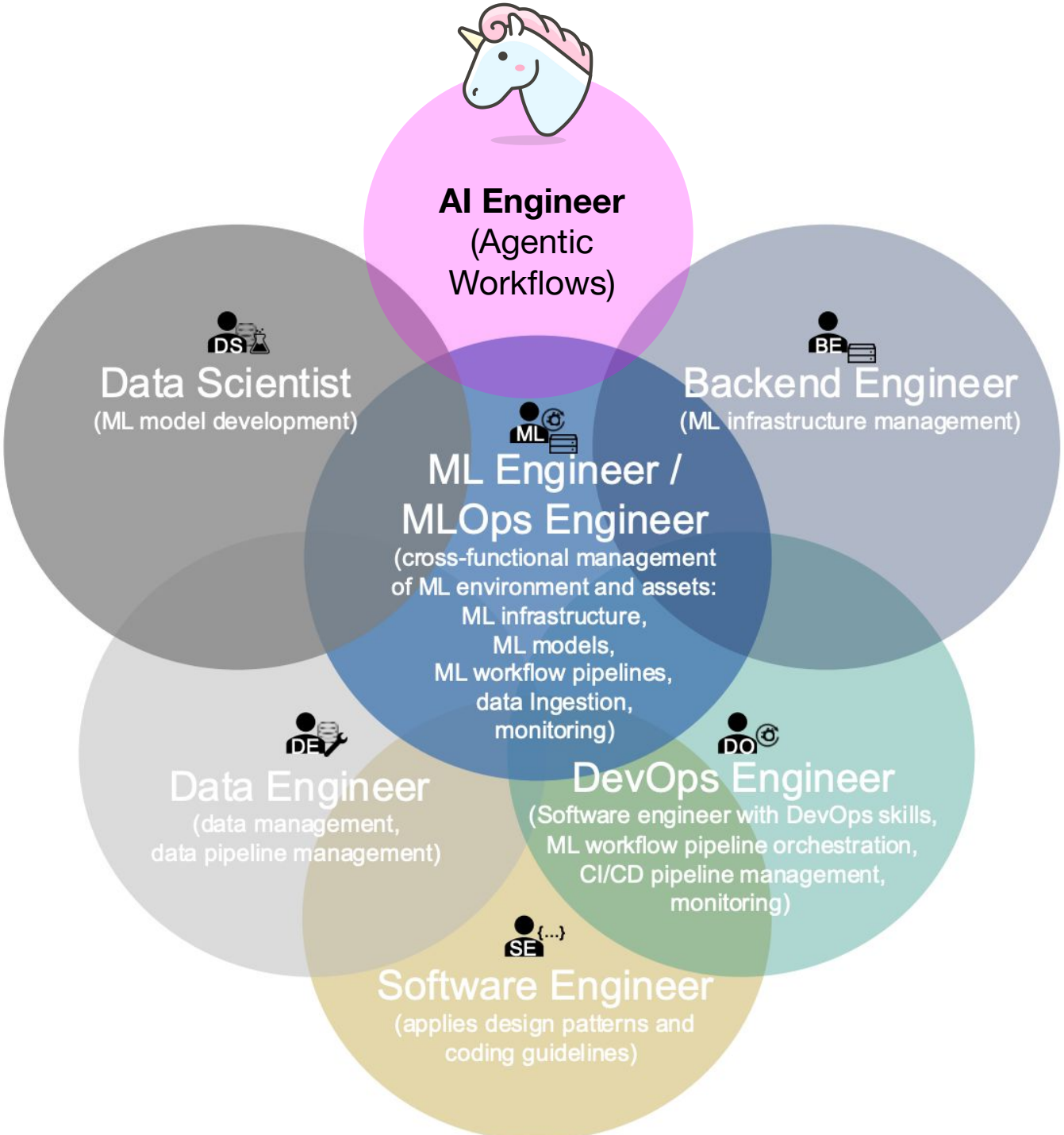
Outlier & Drift Detection

Organisational Process
From SDLC to LLM-DLC

Identifying the **personas** throughout the different **stages** of machine learning development and operation.

Defining the organisation-wide **architectural blueprint** for production machine learning operations.

Bringing together personas, blueprint and processes through the **Development Lifecycle** for large scale **agentic systems**

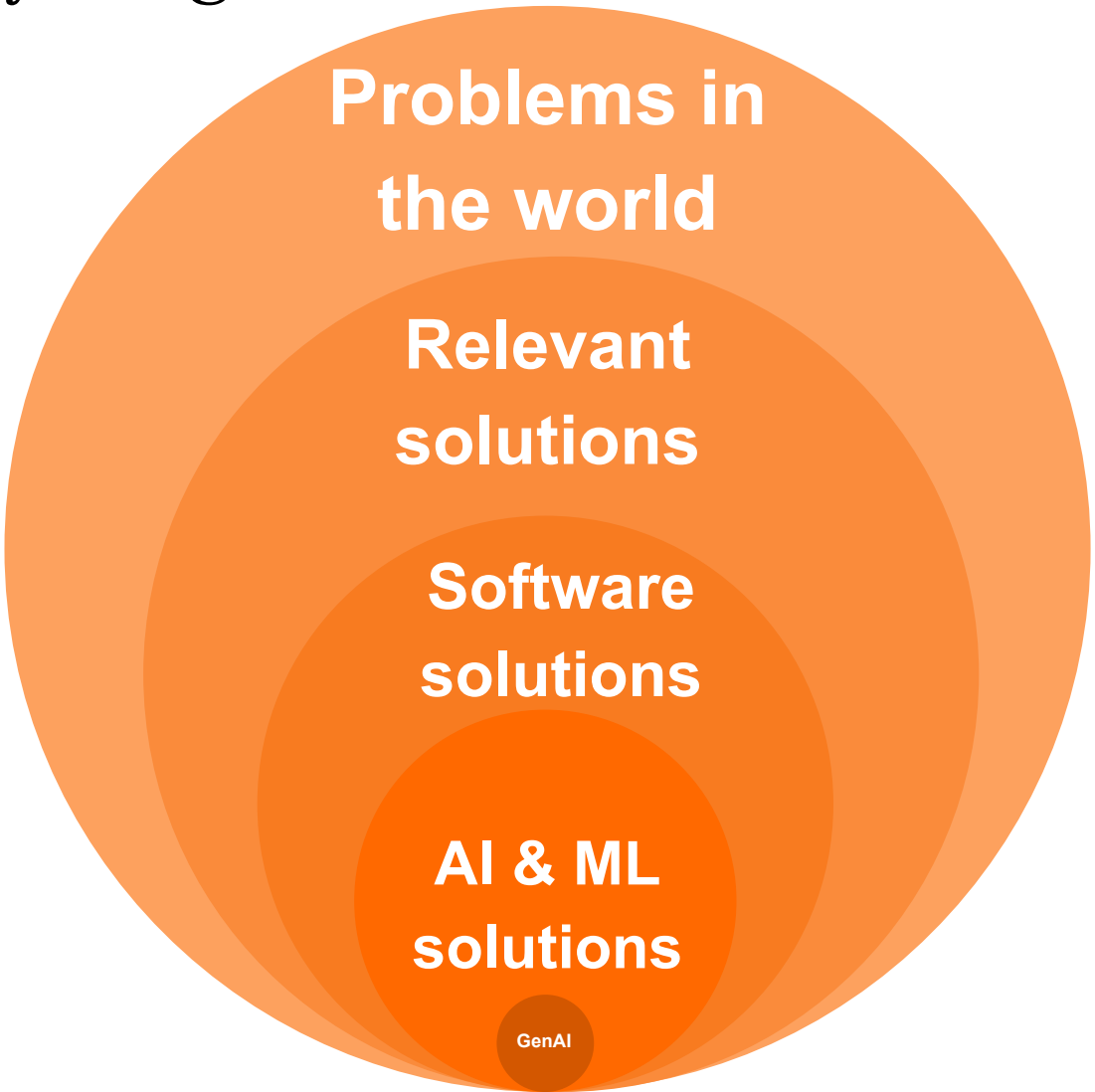


The State of GenAI & ML in Cloud Native Ecosystem

Closing Thoughts

Key Areas to Highlight and Resources

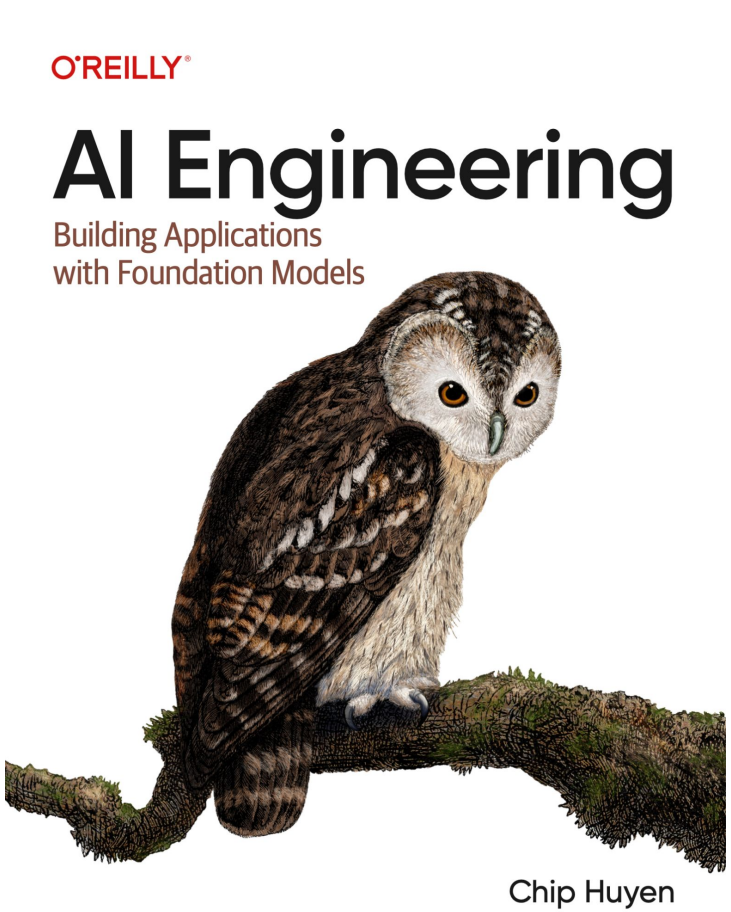
The Reality
Not Everything is AI



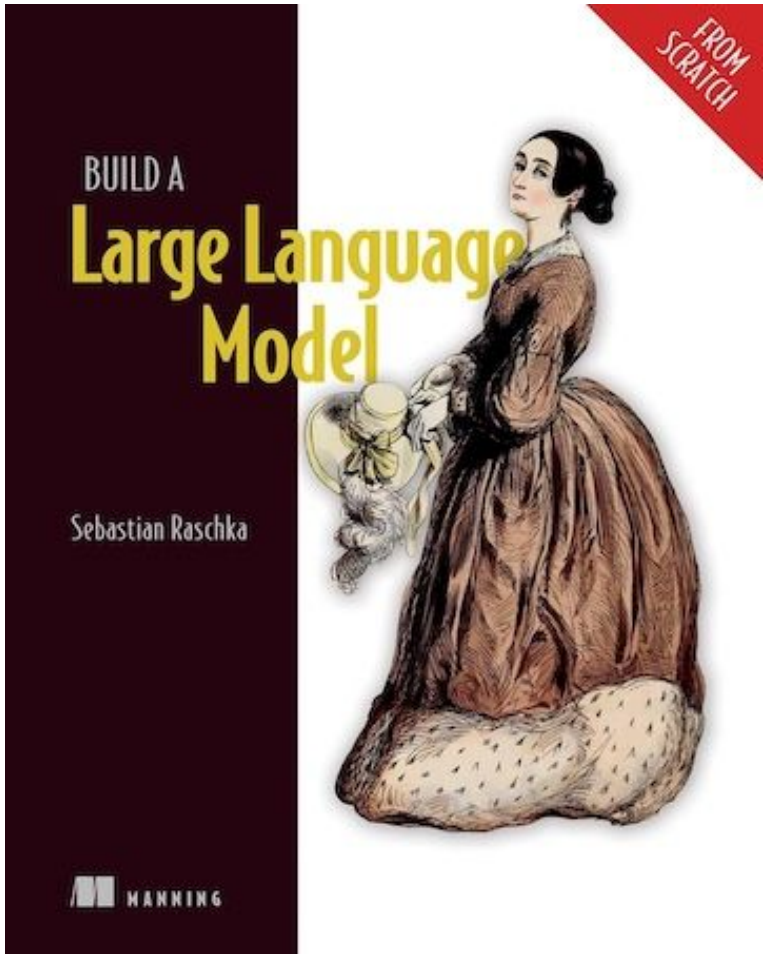
When you run around with a hammer everything may look like a nail

bit.ly/stategenai

Helpful resources
Deep Dives



https://huyenchip.com/books/



https://sebastianraschka.com/books/

The State of GenAI & ML in the Cloud Native Ecosystem

Alejandro Saucedo
Bartosz Ocytko

zalando

Challenges & Patterns
per agency level

Agency	Type	Challenges	Patterns
☆☆☆	**Simple processor**	Hallucinations, latency, model access, prompt quality, prompt injection, data poisoning	Evals, Model proxies, Prompt refinement
★☆☆	**Router**	Pre-defined, static workflows	CoT, ReAct, ADaPT, Reflexion
★★☆	**Tool caller**	Access to diverse tool libraries, wrong tool, params, invalid output	Structured Outputs, MCP (Model Context Protocol)
★★★	**Multi-step agents**	(non-) predictability, planning failures, communication, large prompts	workflow orchestration, visual debuggers, memory, pub/sub, channels, rendezvous (Ada ‘80s)
★★★	**Multi-agent, autonomous agents**	security, rogue agents, endless loops	watchdog, sandbox

https://huggingface.co/docs/smolagents/en/conceptual_guides/intro_agents (agency and type)