AUMOVIO

AURA AI@Production

18. September 2025

Daniel Korom

Location IT Manager

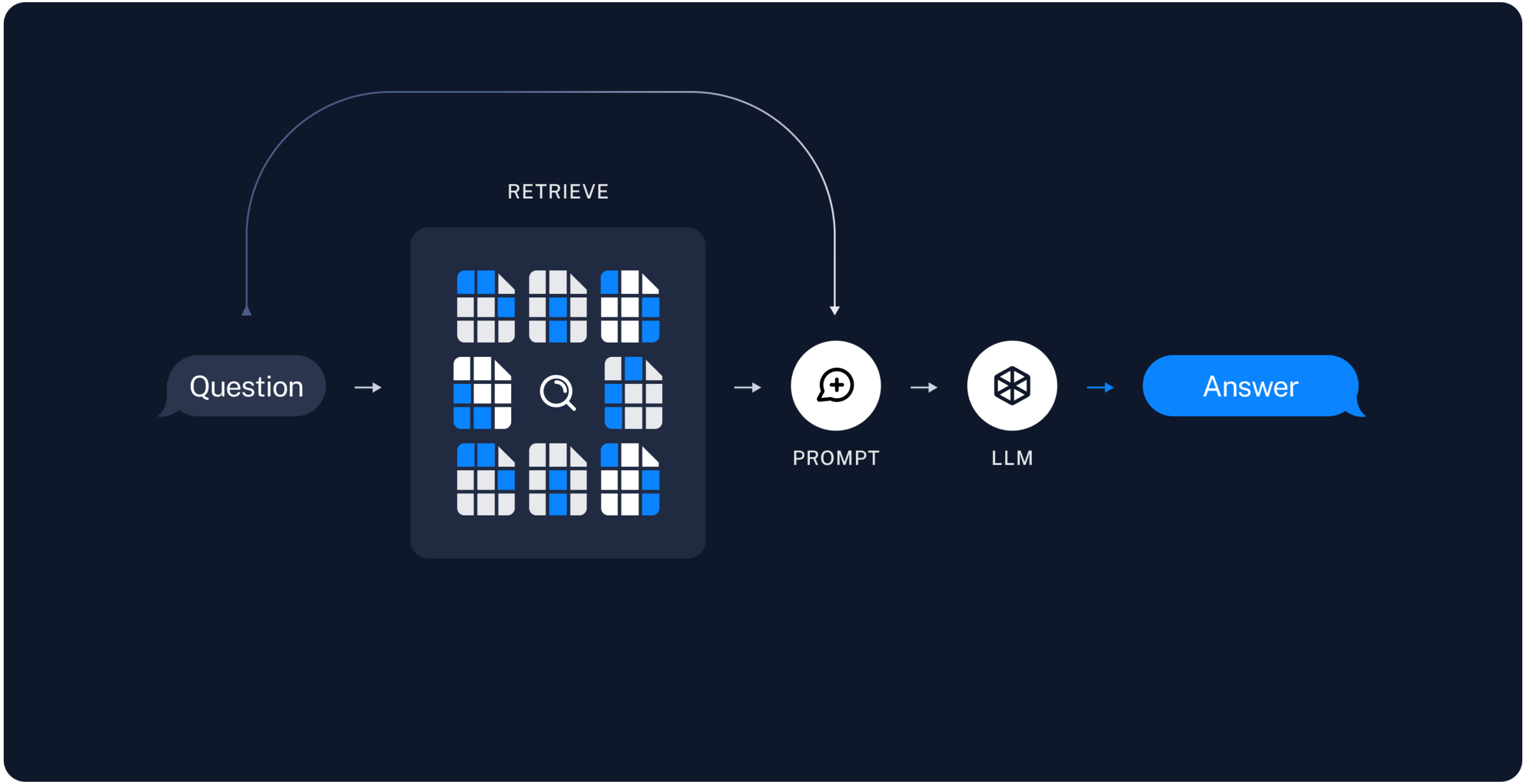
AI@Production
Scope

**We need an assistant,
which helps technicians solve production issues.**

AI@Production
High level overview

Fine-tune LLM	RAG
Take a Large Language Model. Input your texts approximately a million times. Fine-tune the internal weights.	Take a Large Language Model. Create an independent search system for your corpus. Search relevant texts for the question, supply it to the LLM.
+ Quick and easy to run - Really difficult to teach - Hard to control the documents	+ Can be integrated from open-source blocks - Will only use the best-matching documents to generate an answer

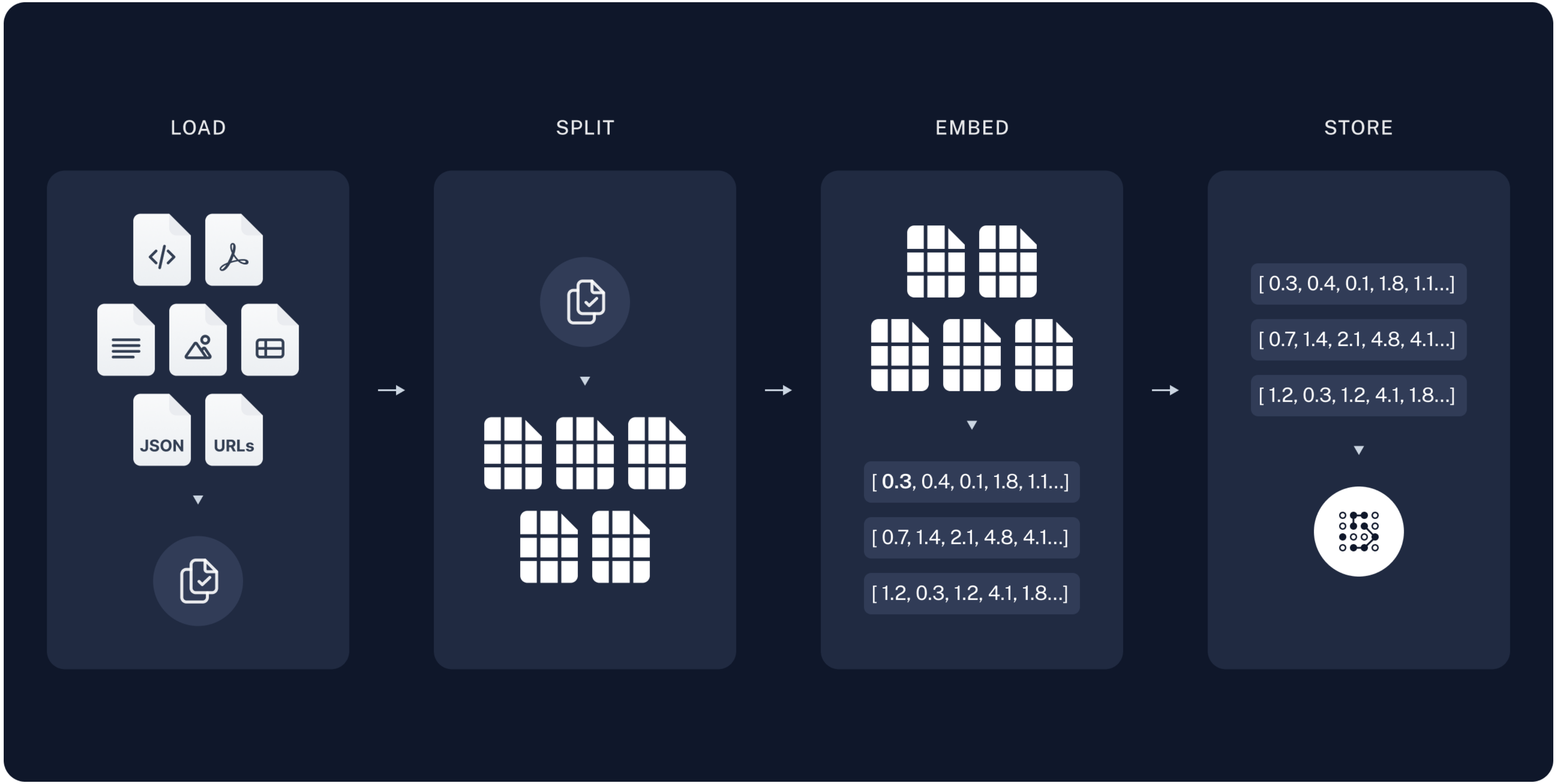
AI@Production
RAG (Retrieval Augmented Generation)



Source: langchain.com

AUMOVIO

AI@Production
RAG (Retrieval Augmented Generation)



Source: langchain.com

AI@Production
Building blocks – cloud/open source

AWS Bedrock
Antropic Claude LLM

Unstructured
Text processing

Streamlit
UI

Langchain, Langgraph, Python
glue for integration

Ollama
Embedding model hosting

BAAI – BGE-M3
Embedding model

Postgres with pgvector
Vector store

AI@Production
Agentic approach - RAG

Provide tools to the LLM and let it decide (based on our written guidance) how to behave.
Easiest option with 2 tools:

- Search the corpus - RAG
- Answer the user

What is the cleaning interval of fixture in M4?

Tool call: context_search{„cleaning interval M4 fixture”}

…

Tool call result: 5x [10 lines] from M4 work instructions

…

Relevant context is enough, ready to answer.

M4 fixture should be cleaned every 8 hours.

AI@Production
Agentic approach - RAG

What will happen if the agent does not have any relevant text?

What is the cleaning interval of fixture in M4?

Tool call: context_search{„cleaning interval M4 fixture”}

...

Tool call result: 5x [10 lines] from irrelevant work instructions

...

Relevant context is not enough, new search.

...

Tool call: context_search{„cleaning M4”}

...

Tool call result: 5x [10 lines] from unrelevant work instructions

There is no information about M4 fixture cleaning.

AI@Production
Agentic approach – IT Tickets

Tools and behaviours:

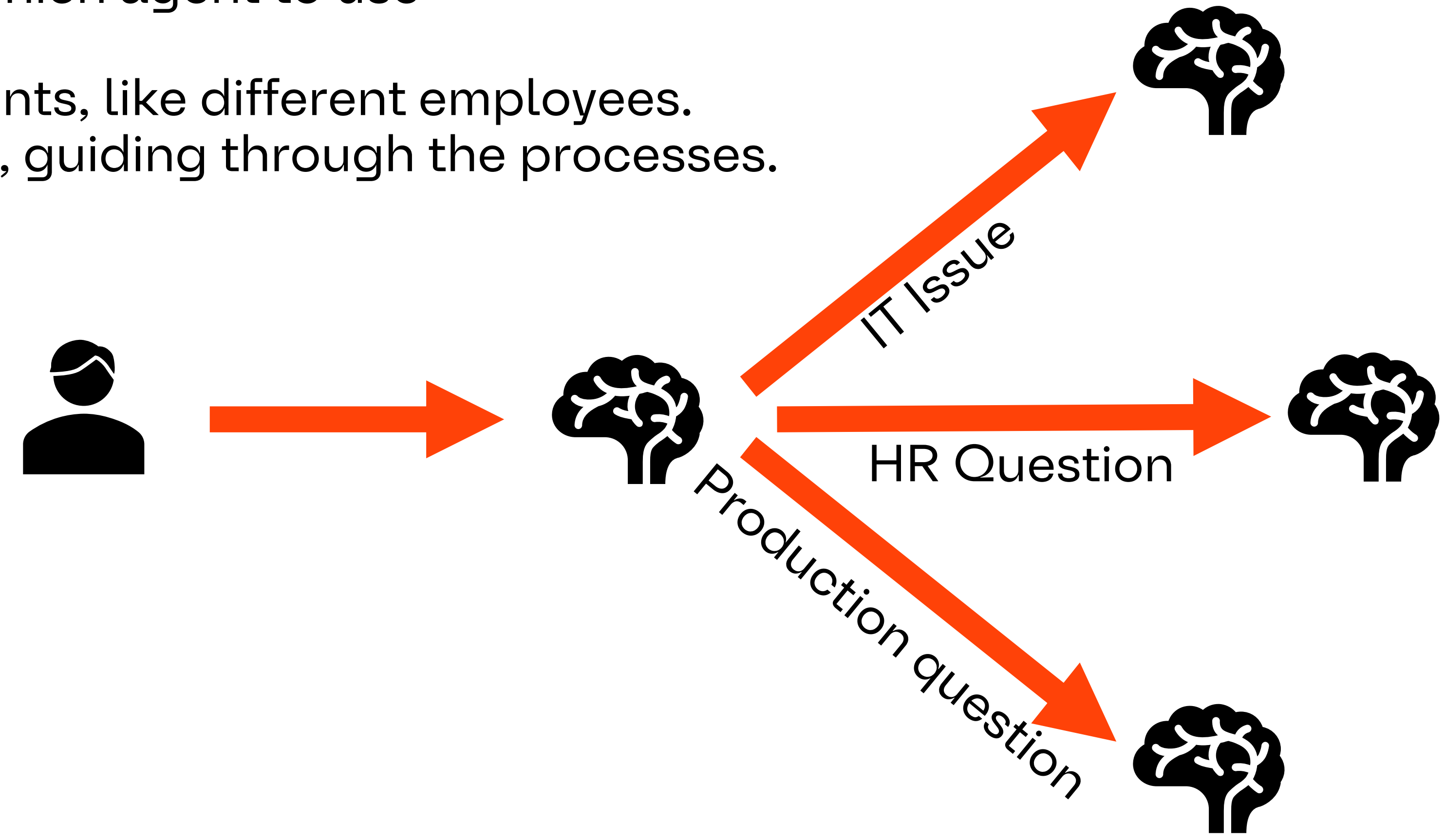
- Anwser (ask questions about poorly problems)
- Search the knowledge base - RAG
- Search the previous tickets and solutions - RAG
- Answer (with a proposed solution)
- Get the devices of the user from the ticketing system
- Create a ticket for the user with the discussed problem
- (only for the brave) Run a script in the background to fix the problem

AI@Production
Agentception

This will lead us to:

- Develop different agents based on business processes
- As this number will increase, it will be hard for users to select
- Supervisory agent to decide which agent to use

We will face a connection of agents, like different employees.
Routing messages to each other, guiding through the processes.



AUMOVIO

AI@Production
Summary

„Quick-win" RAG context search:

- You need relevant text-based, digitalized, documents.
- A process to update these documents, or to remove them from the corpus if not relevant anymore.
- Solutions from the market / smaller-bigger integration and advisory companies.
- With open-source methods ~2 capable and motivated developers in 2 month can have the first PoC.

How to prepare for AI Agents:

- Lean processes are a must before any digitalization – Garbage in → garbage out!
- AI Agents are not always the most beneficial – sometimes a simple internal form is enough. Don't think of this as a one-size-fits-all solution. For human decision, a visualized report is easier to achieve and has greater value.
- LLMs are useful for natural language challenges, if your processes do not need one → use different automation methods.

Thank you!

AUMOVIO