Fsas Technologies
a Fujitsu company

What happens in your data center, stays in your data center.

Zero cloud.
Zero leakage.
Zero excuses.

This is **Private GPT**.
This is **Confidential Mind**.
This is **AI sovereignty**.

Jochen Riedisser

Jochen Riedisser, CDO
ACDC Jam Session
January 15th, 2026

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Our AI buildings block for European Sovereign AI

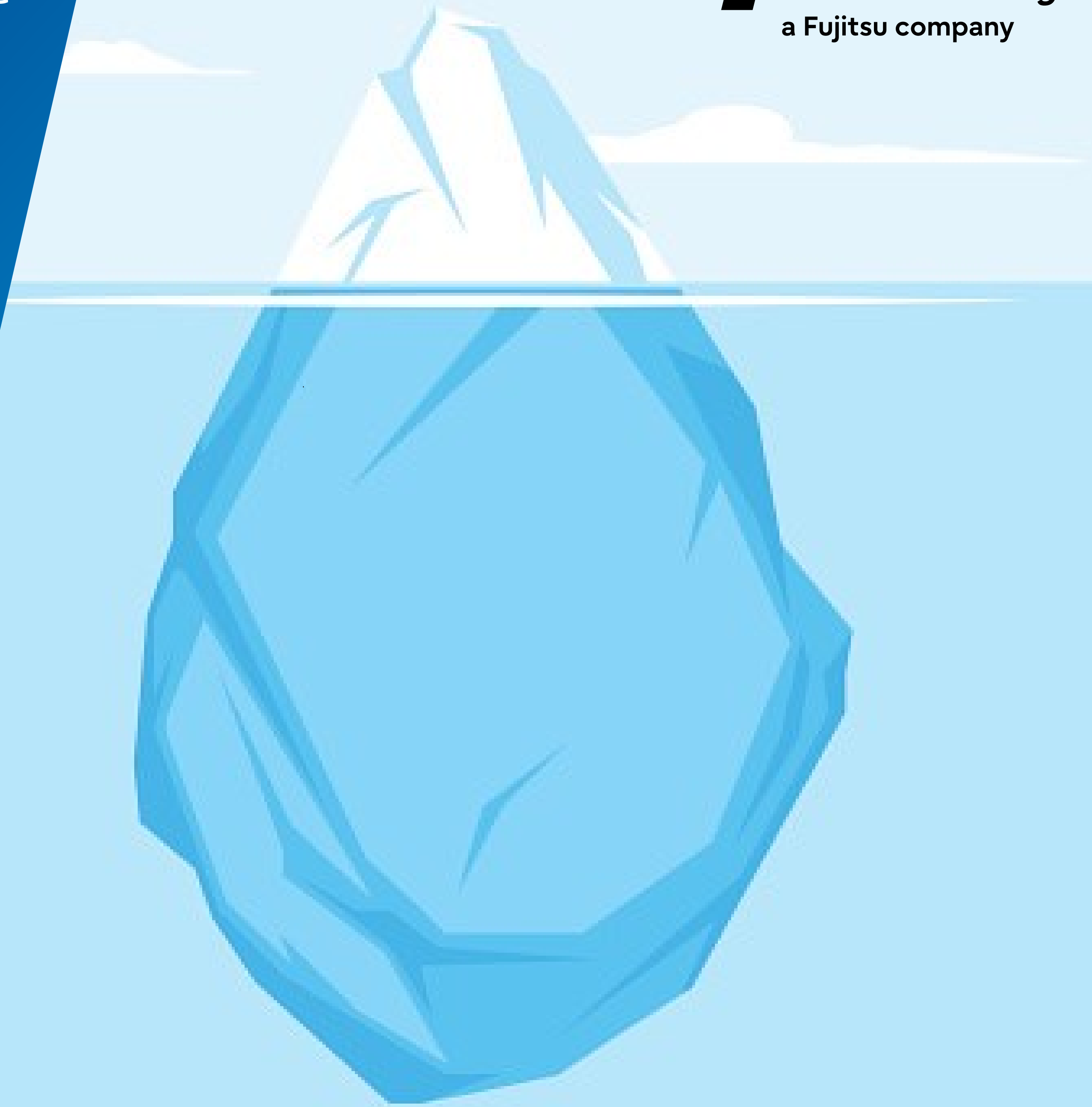
Business	Use Cases
Control & security	Sovereign Data
New skills & knowledge	Trusted AI
Secure, efficient & at scale	Sustainable IT
Regulation	[illegible]

Strong Ecosystem

VNClagoon
SUMM AI
Leichte Sprache. Leicht gemacht.
adesso
elasticsearch
Milvus
Confluence
IONOS
NORTAL
DASU
Transferzentrum für Digitalisierung, Analytics & Data Science Ulm
SOLUTIONS VAGO
Hugging Face
cohere
MISTRAL AI_
universität ulm
NTT
vLLM
RANCHER
NVIDIA
JUNIPER NETWORKS
SUPERMICRO
SUSE
intel
AMD
NetApp

Productive autonomous AI agents at scale

AI iceberg



Applications

Workflow integration

Compliance & governance

Context engineering

Data pipeline

AI workflows (train, build, run)

Security & Monitoring & Orchestration

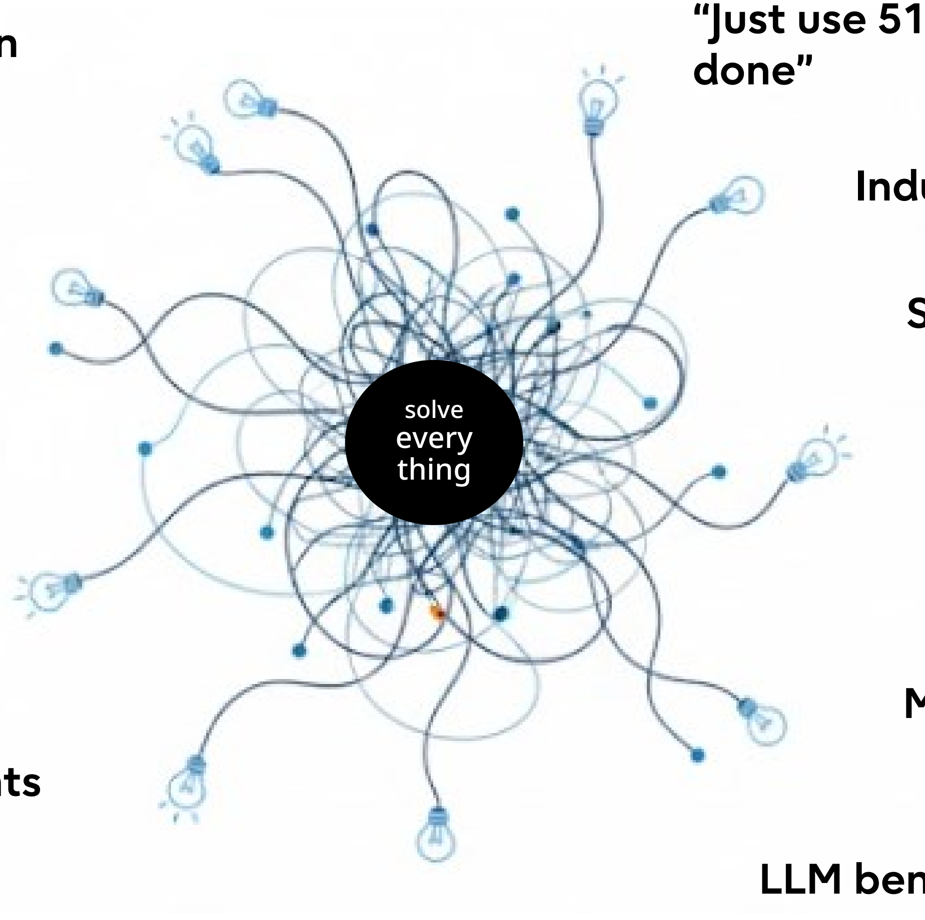
LLM serving & API integration

Kubernetes / Container

AI-ready storage

GPU server

We look after the real challenges



Automation

Tables & Diagrams

Reliable AI endpoints

“We have some databases”

Just ask 8000 documents with a single prompt

Air-gapped AI

“Just use 512-token chunks, done”

Industry data

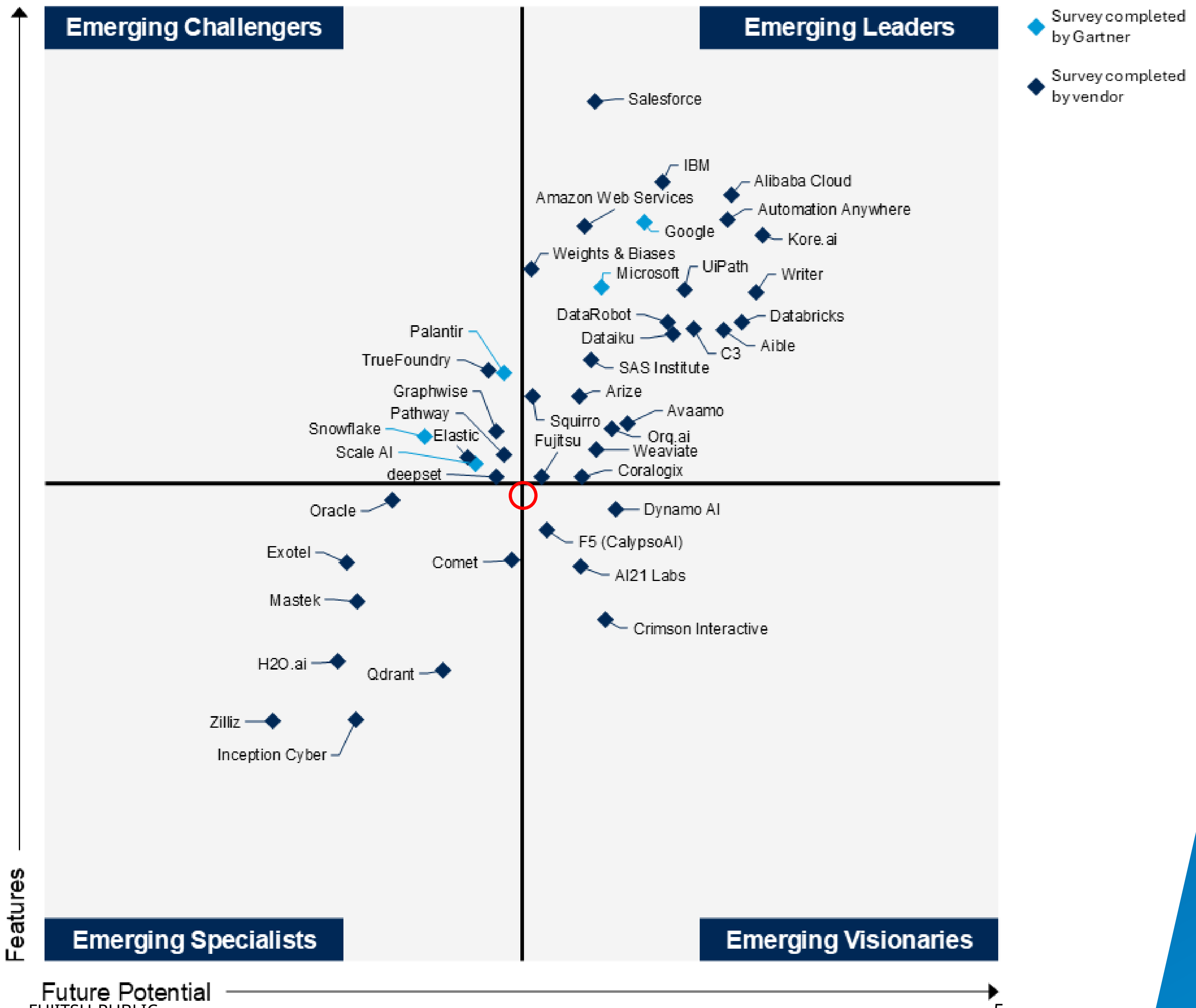
Search engine chaos

Enterprise support on AI

Metadata > Embeddings

LLM benchmark vs reality

Generative AI Engineering



Source: Gartner, as of November 2025 (Gartner aims to update this Emerging Market Quadrant on a quarterly basis on gartner.com)

Pilot

We are leading Generative AI-Engineers.

We are a trusted partner

© Fsas Technologies 2025

Full Stack AI solution provider

Fsas Technologies
a Fujitsu company

On-premise Sovereign AI – European by Design – People you Trust

Build Sovereign Agents together

Customized Agentic AI & Automation DevOps

Langflow
Model Context Protocol
OI Open WebUI
n8n

AI Test Drive

Leading **Data-sovereign AI Ecosystem of Partners**

AI Productivity
Agents | Use Cases

- Knowledge Bot
- Manufacturing
- Digital Admin & IT Services & Coding
- IoT
- Agent AI-Assistant
- Legal & HR
- Retail
- Healthcare

PrivateGPT

Integrated turnkey on-premise AI solution

AI & Data Platform

CONFIDENTIAL Mind
Hugging Face
cohere
milvus
MISTRAL AI_
universität uulm
+
PostgreSQL
vLLM
elasticsearch

Hybrid Cloud Platform

RANCHER

+

AI Infrastructure

NVIDIA
NetApp
SUPERMICRO
intel
JUNIPER NETWORKS

© 2025 Fsas Technologies

Fsas Technologies
a Fujitsu company

Private GPT
the Fsas On-Prem GenAI solution

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Fsas Technologies
a Fujitsu company

Private GPT – an introduction

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Fsas Private GPT Appliance

Fsas Technologies
a Fujitsu company

European
Large
Language
Model

Open Source

+

Your own
data
(RAG)

=

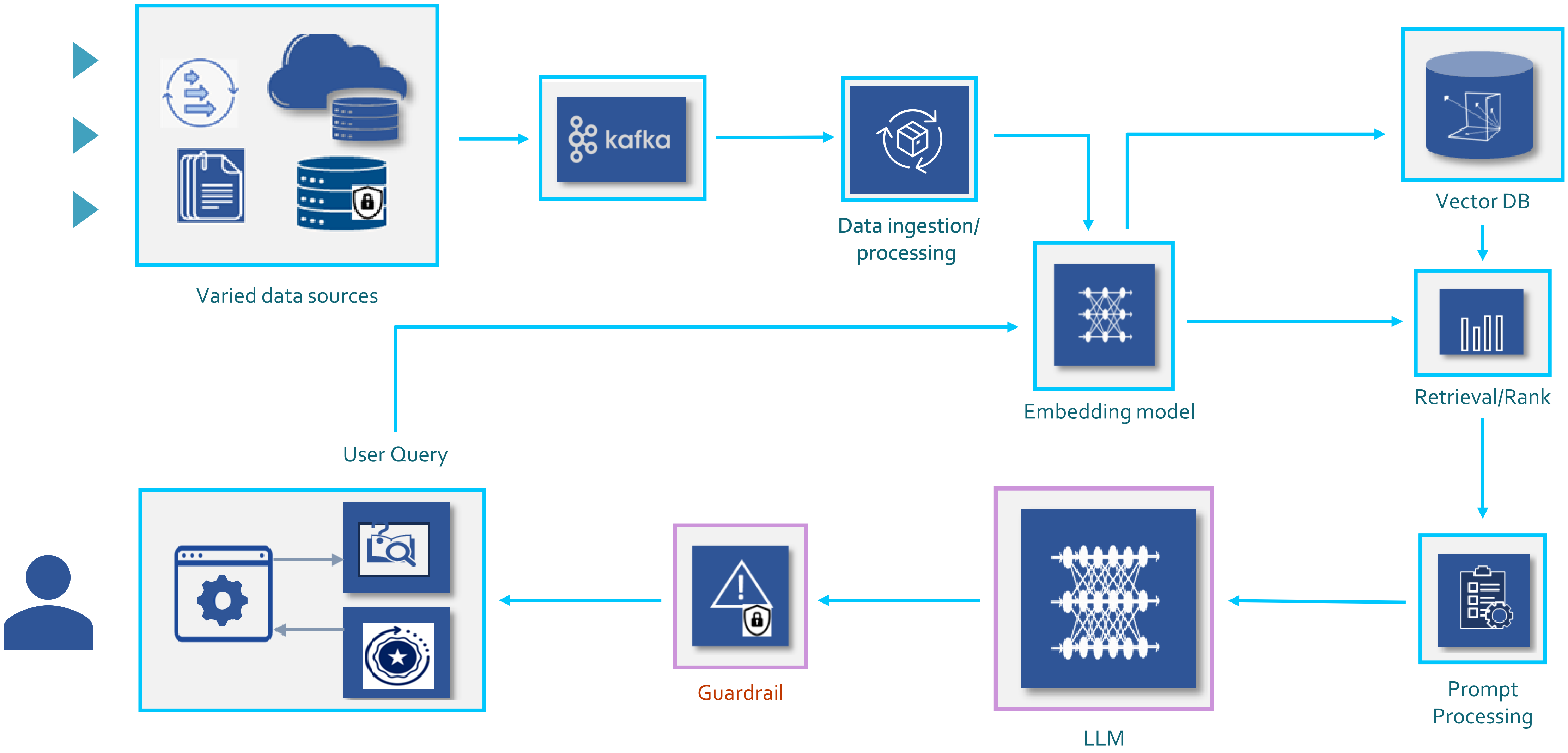
PGPT
use AI and
your own
knowledge
while
keeping it
sovereign
and
protected

Fsas PRIMERGY Server plus AI accelerator

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

How our Retrieval Augmented Generation (RAG) works

Fsas Technologies
a Fujitsu company



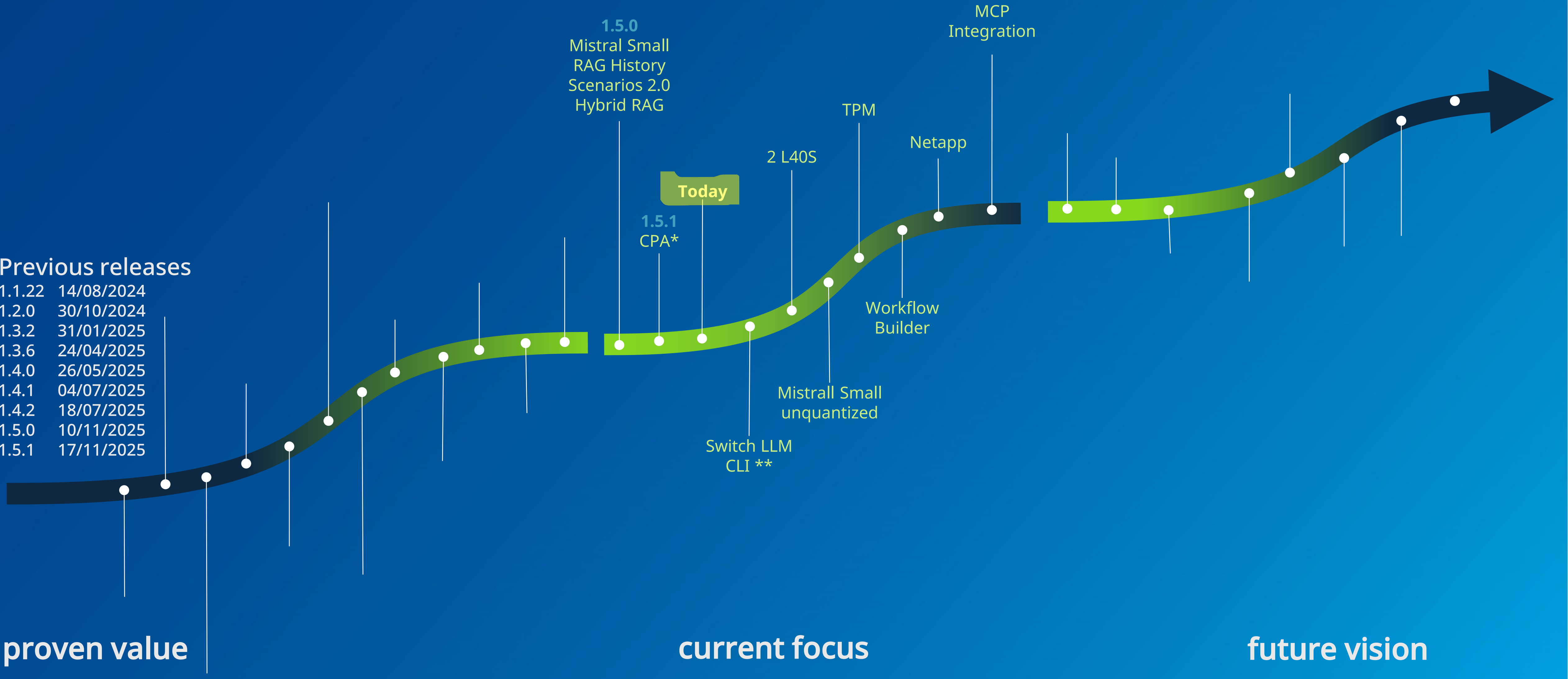
PGPT

Fsas Technologies
a Fujitsu company

Value Stream

Previous releases

1.1.22	14/08/2024
1.2.0	30/10/2024
1.3.2	31/01/2025
1.3.6	24/04/2025
1.4.0	26/05/2025
1.4.1	04/07/2025
1.4.2	18/07/2025
1.5.0	10/11/2025
1.5.1	17/11/2025



1.5.0
Mistral Small
RAG History
Scenarios 2.0
Hybrid RAG

1.5.1
CPA*

Today

Switch LLM
CLI **

2 L40S

Mistrall Small
unquantized

TPM

Workflow
Builder

Netapp

MCP
Integration

proven value

current focus

future vision

DGPT

powered by

Mistral AI

Fsas Technologies
a Fujitsu company

upgrade

Mistral Small 3

1. **Lightweight**:
   Mistral Small 3 can run on a single L40s and newer.
   This makes it a great fit for on-device use cases.
2. **Fast-response conversational assistance**:
   Ideal for virtual assistants and other applications where quick, accurate responses are essential.
3. **Low-latency function calling**:
   Capable of rapid function execution within automated or agentic workflows

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Introducing scenarios

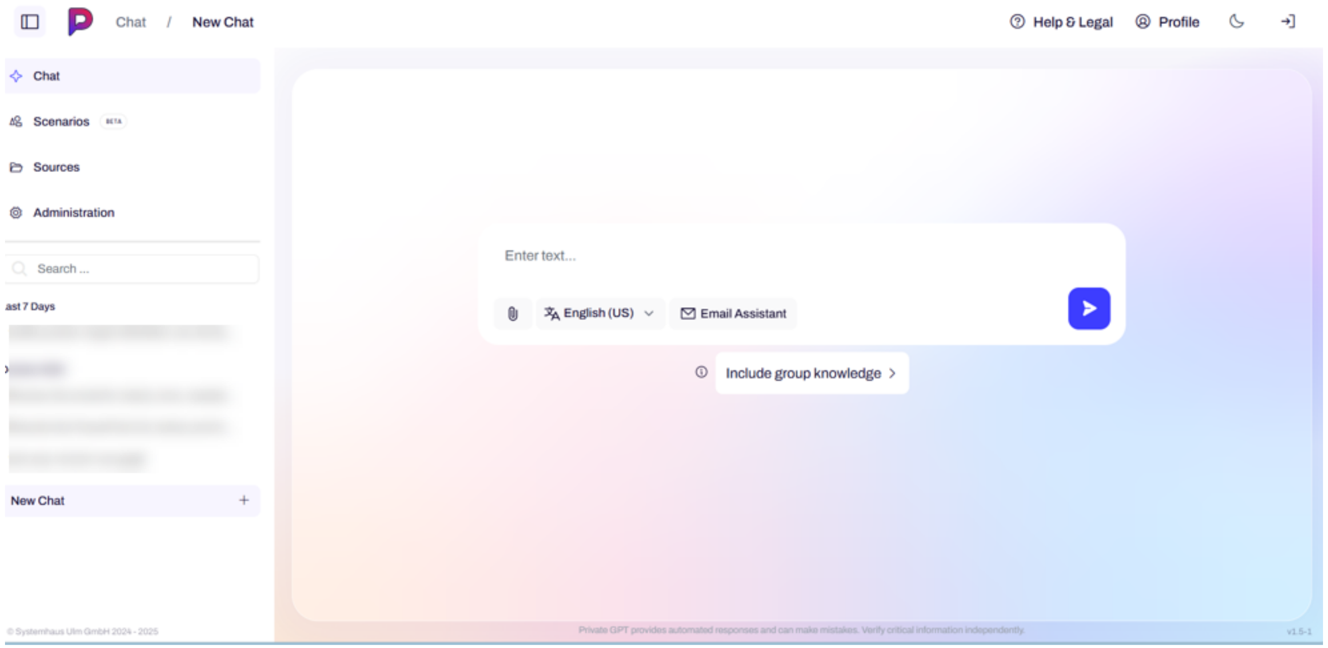
See it as a template for repeating tasks:

Instead of many times repeating the same prompt, you can set it up once and use it many times

There are additional options:

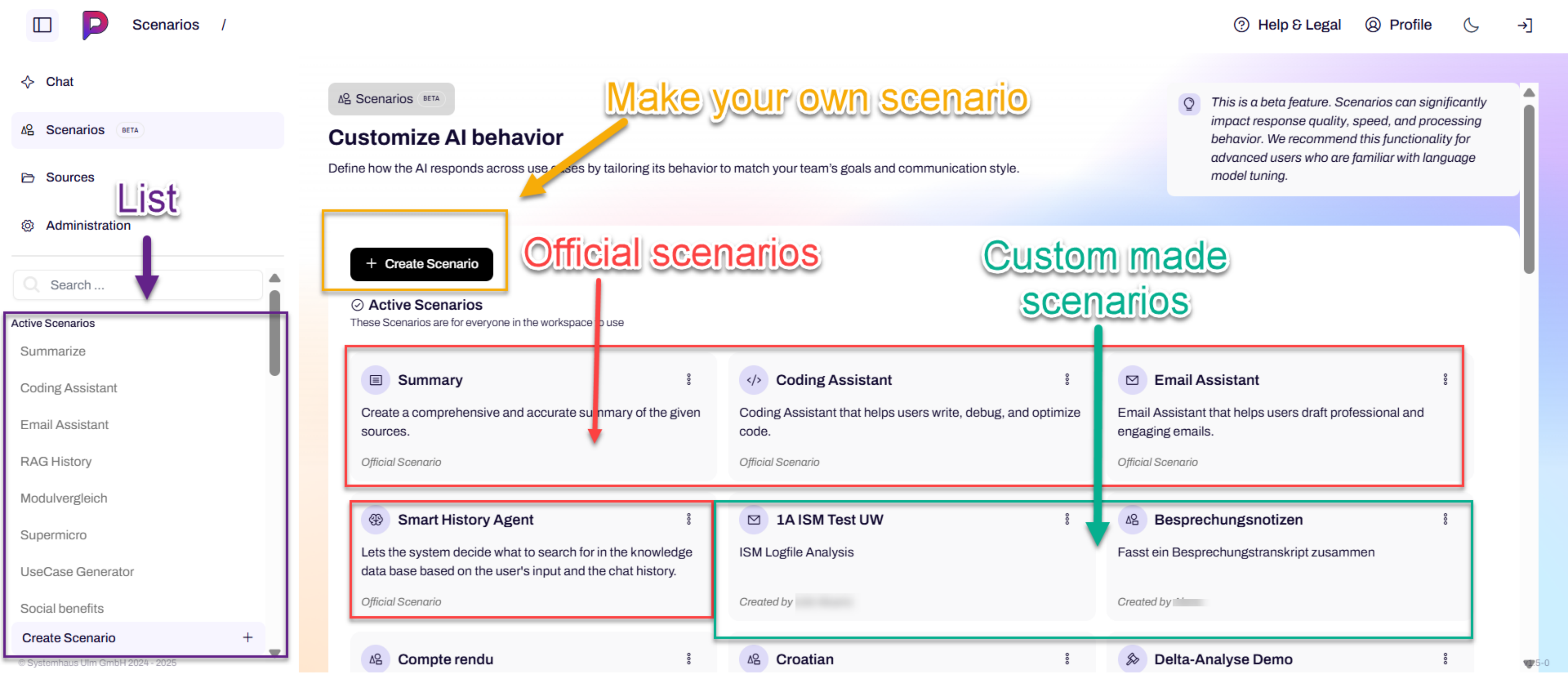
- To use extra information from documents
- To set parameters to influence system behaviour
- To give instructions on the format of the output
- And more

This is what the workshop is about



Private GPT provides automated responses and can make mistakes. Verify critical information independently.

Scenario page



List

Active Scenarios

- Summarize
- Coding Assistant
- Email Assistant
- RAG History
- Modulvergleich
- Supermicro
- UseCase Generator
- Social benefits

Create Scenario +

© Systemhaus Ulm GmbH 2024 - 2025

Scenarios BETA

Make your own scenario

Customize AI behavior

Define how the AI responds across use cases by tailoring its behavior to match your team's goals and communication style.

This is a beta feature. Scenarios can significantly impact response quality, speed, and processing behavior. We recommend this functionality for advanced users who are familiar with language model tuning.

+ Create Scenario

Official scenarios

Custom made scenarios

Active Scenarios

These Scenarios are for everyone in the workspace to use

Summary
Create a comprehensive and accurate summary of the given sources.
Official Scenario

Coding Assistant
Coding Assistant that helps users write, debug, and optimize code.
Official Scenario

Email Assistant
Email Assistant that helps users draft professional and engaging emails.
Official Scenario

Smart History Agent
Lets the system decide what to search for in the knowledge data base based on the user's input and the chat history.
Official Scenario

1A ISM Test UW
ISM Logfile Analysis
Created by [redacted]

Besprechungsnotizen
Fasst ein Besprechungstranskript zusammen
Created by [redacted]

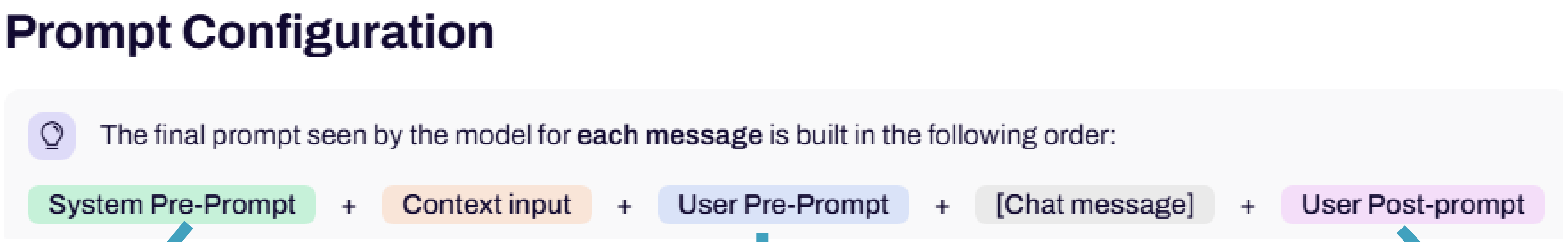
Compte rendu

Croatian

Delta-Analyse Demo

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

The Pre-Prompts



Prompt Configuration

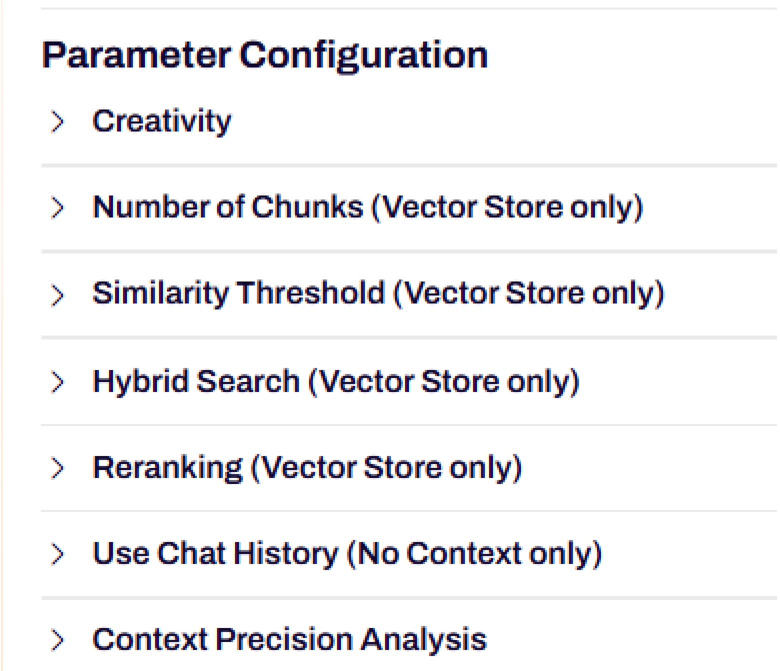
The final prompt seen by the model for **each message** is built in the following order:

System Pre-Prompt + Context input + User Pre-Prompt + [Chat message] + User Post-prompt

System Pre-Prompt:
The Chef's General Training and Kitchen Rules

User Pre-Prompt:
The Customer's Specific Request (Before the Main Order)

User Post-Prompt**:**
The Customer's Final Instructions (After the Main Order)

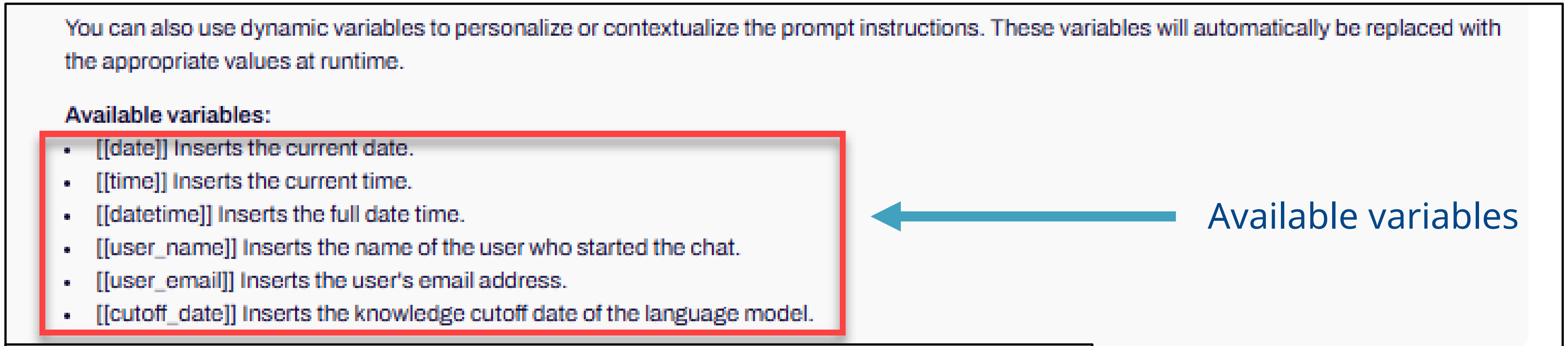


Parameter Configuration

- Creativity
- Number of Chunks (Vector Store only)
- Similarity Threshold (Vector Store only)
- Hybrid Search (Vector Store only)
- Reranking (Vector Store only)
- Use Chat History (No Context only)
- Context Precision Analysis

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Using Variables

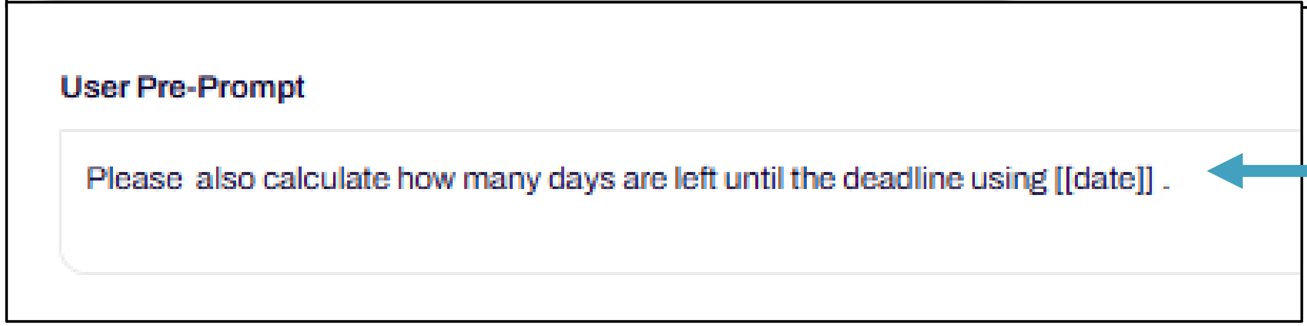


You can also use dynamic variables to personalize or contextualize the prompt instructions. These variables will automatically be replaced with the appropriate values at runtime.

Available variables:

- [[date]] Inserts the current date.
- [[time]] Inserts the current time.
- [[datetime]] Inserts the full date time.
- [[user_name]] Inserts the name of the user who started the chat.
- [[user_email]] Inserts the user's email address.
- [[cutoff_date]] Inserts the knowledge cutoff date of the language model.

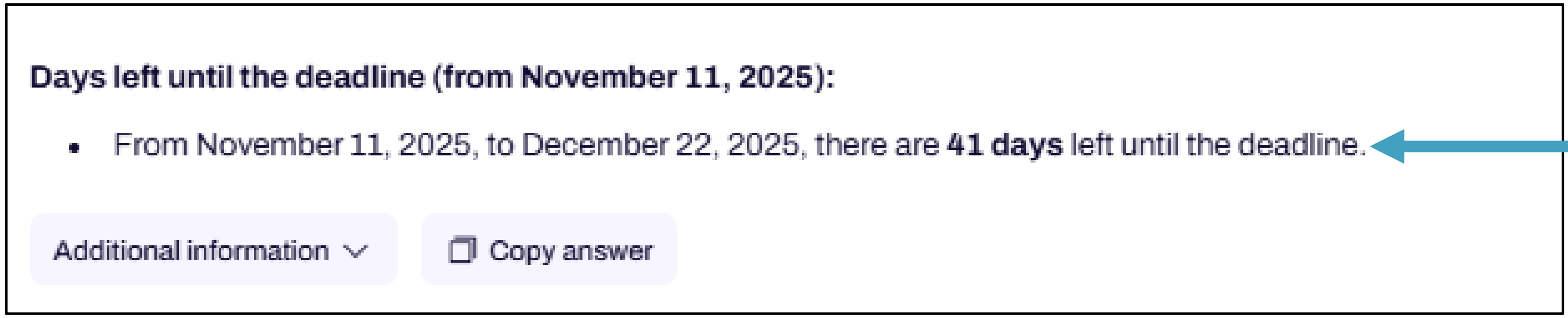
Available variables



User Pre-Prompt

Please also calculate how many days are left until the deadline using [[date]] .

Use in prompt

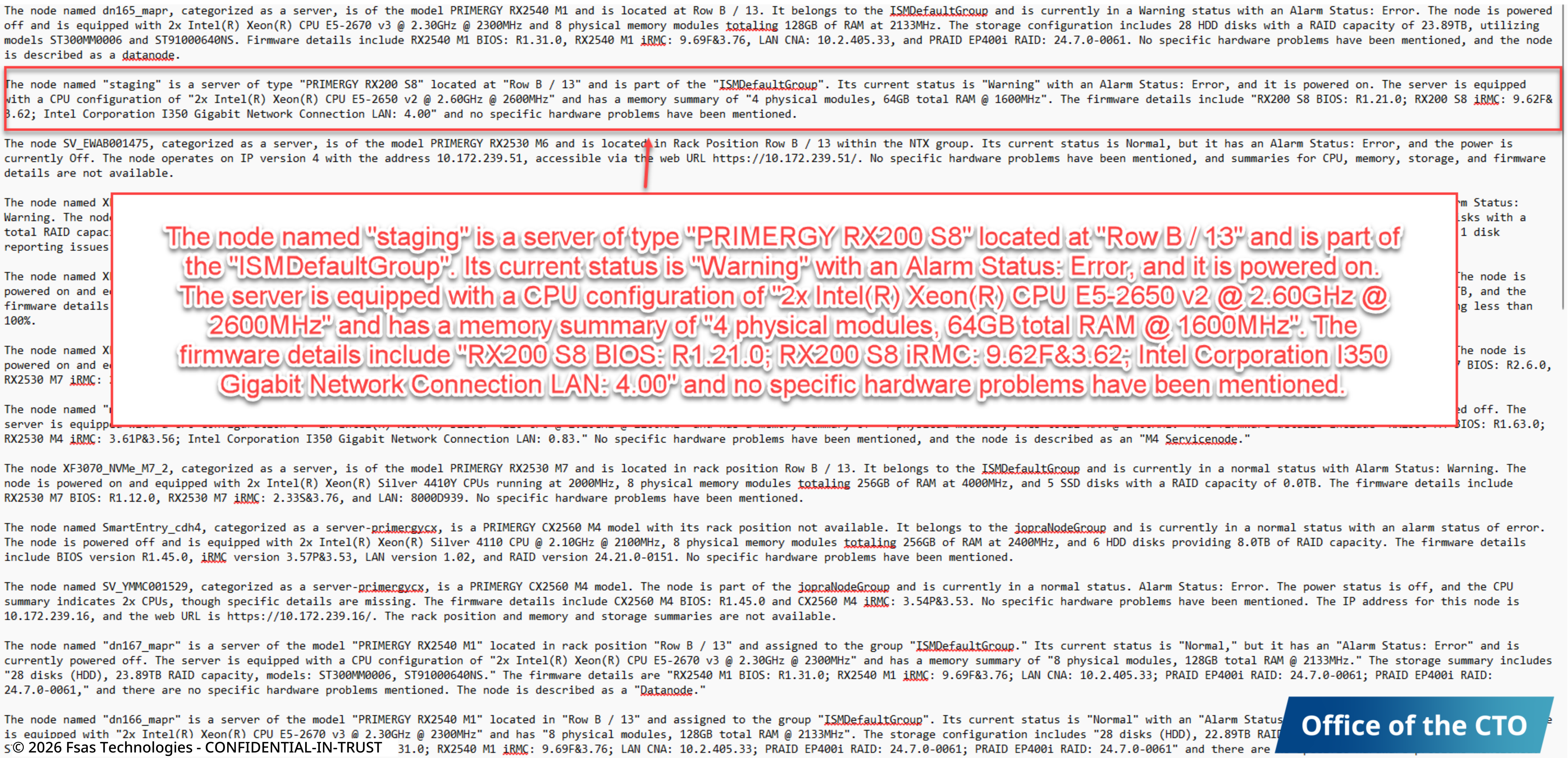


Days left until the deadline (from November 11, 2025):

- From November 11, 2025, to December 22, 2025, there are **41 days** left until the deadline.

Answer when using scenario.
But careful with calculations, they might not be right all the time!

Example: My data



The node named dn165_mapr, categorized as a server, is of the model PRIMERGY RX2540 M1 and is located at Row B / 13. It belongs to the ISMDefaultGroup and is currently in a Warning status with an Alarm Status: Error. The node is powered off and is equipped with 2x Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz @ 2300MHz and 8 physical memory modules totaling 128GB of RAM at 2133MHz. The storage configuration includes 28 HDD disks with a RAID capacity of 23.89TB, utilizing models ST300MM0006 and ST91000640NS. Firmware details include RX2540 M1 BIOS: R1.31.0, RX2540 M1 iRMC: 9.69F&3.76, LAN CNA: 10.2.405.33, and PRAID EP400i RAID: 24.7.0-0061. No specific hardware problems have been mentioned, and the node is described as a datanode.

The node named "staging" is a server of type "PRIMERGY RX200 S8" located at "Row B / 13" and is part of the "ISMDefaultGroup". Its current status is "Warning" with an Alarm Status: Error, and it is powered on. The server is equipped with a CPU configuration of "2x Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz @ 2600MHz" and has a memory summary of "4 physical modules, 64GB total RAM @ 1600MHz". The firmware details include "RX200 S8 BIOS: R1.21.0; RX200 S8 iRMC: 9.62F&3.62; Intel Corporation I350 Gigabit Network Connection LAN: 4.00" and no specific hardware problems have been mentioned.

The node SV_EWAB001475, categorized as a server, is of the model PRIMERGY RX2530 M6 and is located in Rack Position Row B / 13 within the NTX group. Its current status is Normal, but it has an Alarm Status: Error, and the power is currently Off. The node operates on IP version 4 with the address 10.172.239.51, accessible via the web URL https://10.172.239.51/. No specific hardware problems have been mentioned, and summaries for CPU, memory, storage, and firmware details are not available.

The node named X m Status:
Warning. The nod sks with a
total RAID capac 1 disk
reporting issues

The node named X he node is
powered on and e B, and the
firmware details g less than
100%.

The node named X he node is
powered on and e BIOS: R2.6.0,
RX2530 M7 iRMC:

The node named " d off. The
server is equipp BIOS: R1.63.0;
RX2530 M4 iRMC: 3.61P&3.56; Intel Corporation I350 Gigabit Network Connection LAN: 0.83." No specific hardware problems have been mentioned, and the node is described as an "M4 Servicenode."

The node XF3070_NVMe_M7_2, categorized as a server, is of the model PRIMERGY RX2530 M7 and is located in rack position Row B / 13. It belongs to the ISMDefaultGroup and is currently in a normal status with Alarm Status: Warning. The node is powered on and equipped with 2x Intel(R) Xeon(R) Silver 4410Y CPUs running at 2000MHz, 8 physical memory modules totaling 256GB of RAM at 4000MHz, and 5 SSD disks with a RAID capacity of 0.0TB. The firmware details include RX2530 M7 BIOS: R1.12.0, RX2530 M7 iRMC: 2.33S&3.76, and LAN: 8000D939. No specific hardware problems have been mentioned.

The node named SmartEntry_cdh4, categorized as a server-primergycx, is a PRIMERGY CX2560 M4 model with its rack position not available. It belongs to the jopraNodeGroup and is currently in a normal status with an alarm status of error. The node is powered off and is equipped with 2x Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz @ 2100MHz, 8 physical memory modules totaling 256GB of RAM at 2400MHz, and 6 HDD disks providing 8.0TB of RAID capacity. The firmware details include BIOS version R1.45.0, iRMC version 3.57P&3.53, LAN version 1.02, and RAID version 24.21.0-0151. No specific hardware problems have been mentioned.

The node named SV_YMMC001529, categorized as a server-primergycx, is a PRIMERGY CX2560 M4 model. The node is part of the jopraNodeGroup and is currently in a normal status. Alarm Status: Error. The power status is off, and the CPU summary indicates 2x CPUs, though specific details are missing. The firmware details include CX2560 M4 BIOS: R1.45.0 and CX2560 M4 iRMC: 3.54P&3.53. No specific hardware problems have been mentioned. The IP address for this node is 10.172.239.16, and the web URL is https://10.172.239.16/. The rack position and memory and storage summaries are not available.

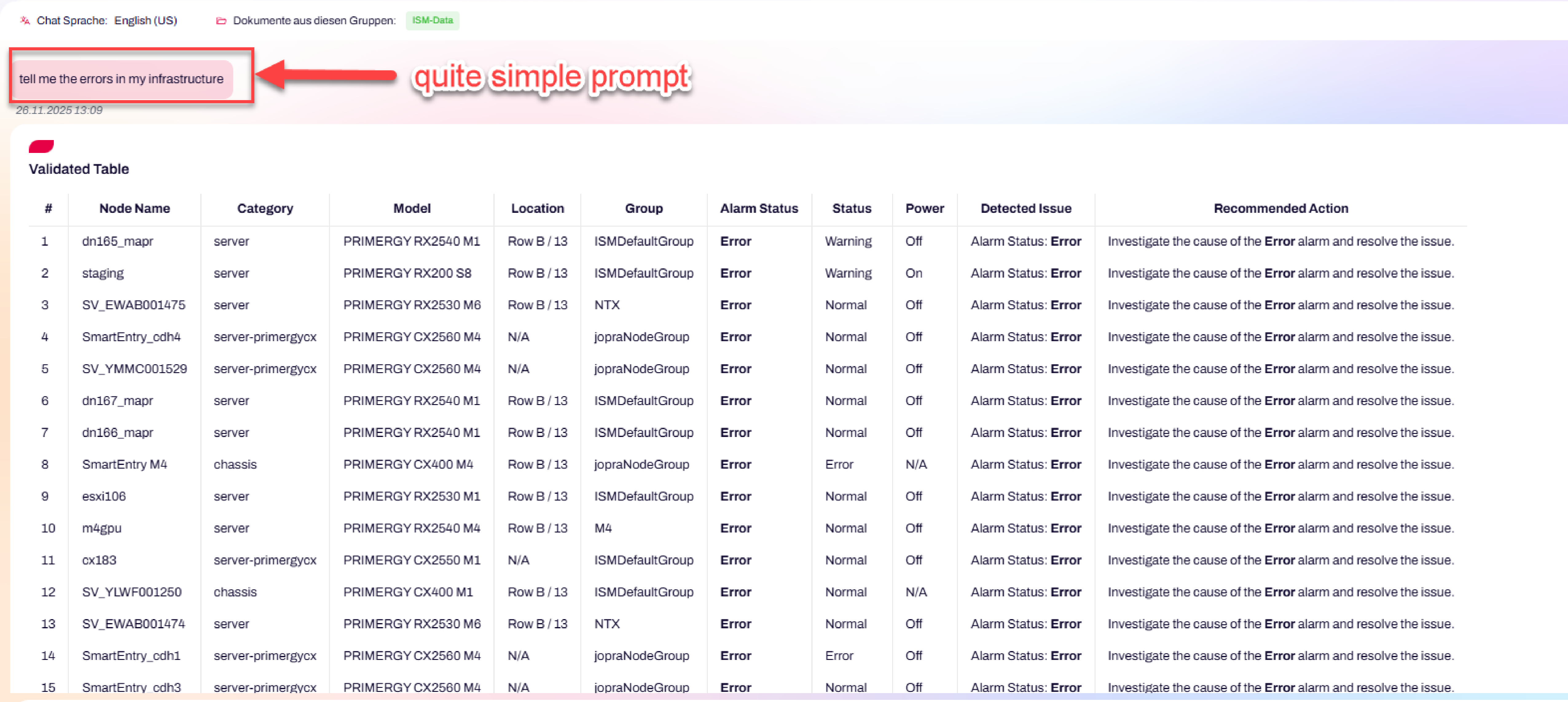
The node named "dn167_mapr" is a server of the model "PRIMERGY RX2540 M1" located in rack position "Row B / 13" and assigned to the group "ISMDefaultGroup." Its current status is "Normal," but it has an "Alarm Status: Error" and is currently powered off. The server is equipped with a CPU configuration of "2x Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz @ 2300MHz" and has a memory summary of "8 physical modules, 128GB total RAM @ 2133MHz." The storage summary includes "28 disks (HDD), 23.89TB RAID capacity, models: ST300MM0006, ST91000640NS." The firmware details are "RX2540 M1 BIOS: R1.31.0; RX2540 M1 iRMC: 9.69F&3.76; LAN CNA: 10.2.405.33; PRAID EP400i RAID: 24.7.0-0061; PRAID EP400i RAID: 24.7.0-0061," and there are no specific hardware problems mentioned. The node is described as a "Datanode."

The node named "dn166_mapr" is a server of the model "PRIMERGY RX2540 M1" located in "Row B / 13" and assigned to the group "ISMDefaultGroup". Its current status is "Normal" with an "Alarm Status
is equipped with "2x Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz @ 2300MHz" and has "8 physical modules, 128GB total RAM @ 2133MHz". The storage configuration includes "28 disks (HDD), 22.89TB RAI
31.0; RX2540 M1 iRMC: 9.69F&3.76; LAN CNA: 10.2.405.33; PRAID EP400i RAID: 24.7.0-0061; PRAID EP400i RAID: 24.7.0-0061" and there are

The node named "staging" is a server of type "PRIMERGY RX200 S8" located at "Row B / 13" and is part of the "ISMDefaultGroup". Its current status is "Warning" with an Alarm Status: Error, and it is powered on. The server is equipped with a CPU configuration of "2x Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz @ 2600MHz" and has a memory summary of "4 physical modules, 64GB total RAM @ 1600MHz". The firmware details include "RX200 S8 BIOS: R1.21.0; RX200 S8 iRMC: 9.62F&3.62; Intel Corporation I350 Gigabit Network Connection LAN: 4.00" and no specific hardware problems have been mentioned.

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Example: My outcome



Chat Sprache: English (US)
Dokumente aus diesen Gruppen: ISM-Data

tell me the errors in my infrastructure

quite simple prompt

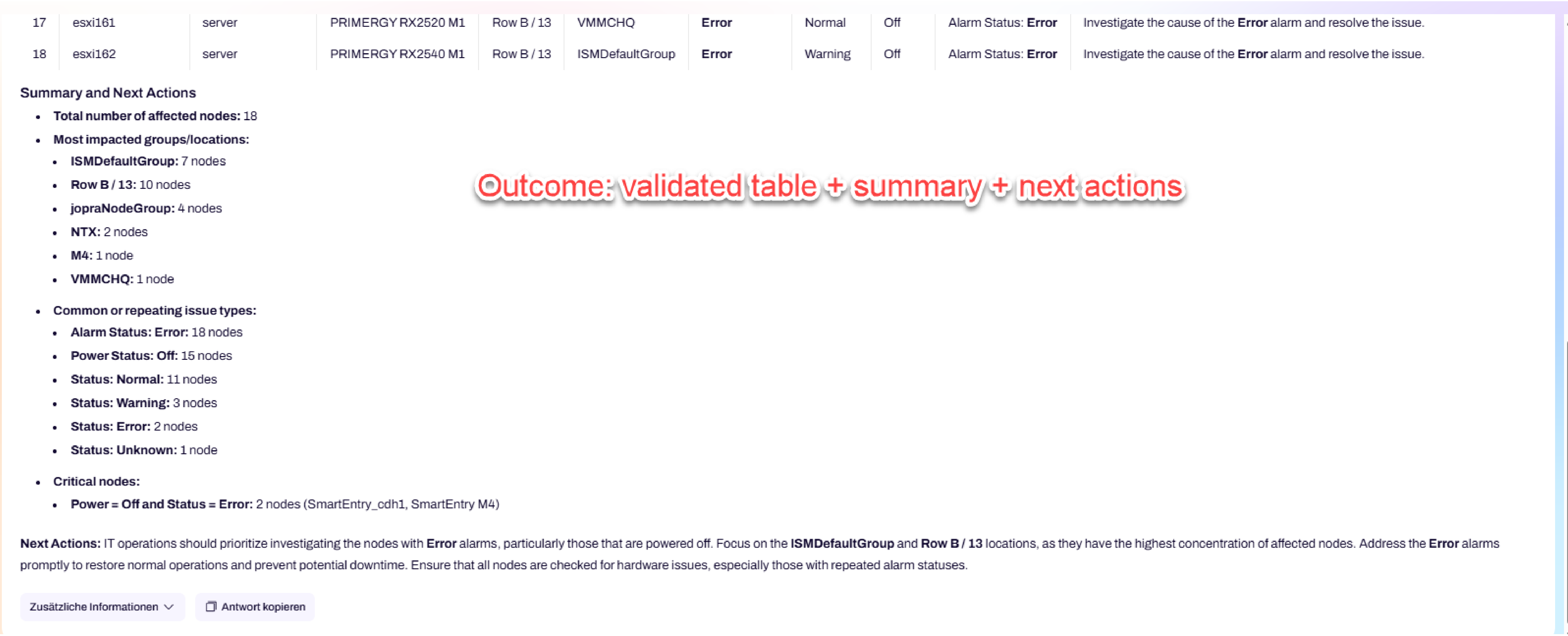
26.11.2025 13:09

Validated Table

#	Node Name	Category	Model	Location	Group	Alarm Status	Status	Power	Detected Issue	Recommended Action
1	dn165_mapr	server	PRIMERGY RX2540 M1	Row B / 13	ISMDefaultGroup	**Error**	Warning	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
2	staging	server	PRIMERGY RX200 S8	Row B / 13	ISMDefaultGroup	**Error**	Warning	On	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
3	SV_EWAB001475	server	PRIMERGY RX2530 M6	Row B / 13	NTX	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
4	SmartEntry_cdh4	server-primergycx	PRIMERGY CX2560 M4	N/A	jopraNodeGroup	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
5	SV_YMMC001529	server-primergycx	PRIMERGY CX2560 M4	N/A	jopraNodeGroup	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
6	dn167_mapr	server	PRIMERGY RX2540 M1	Row B / 13	ISMDefaultGroup	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
7	dn166_mapr	server	PRIMERGY RX2540 M1	Row B / 13	ISMDefaultGroup	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
8	SmartEntry M4	chassis	PRIMERGY CX400 M4	Row B / 13	jopraNodeGroup	**Error**	Error	N/A	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
9	esxi106	server	PRIMERGY RX2530 M1	Row B / 13	ISMDefaultGroup	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
10	m4gpu	server	PRIMERGY RX2540 M4	Row B / 13	M4	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
11	cx183	server-primergycx	PRIMERGY CX2550 M1	N/A	ISMDefaultGroup	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
12	SV_YLWF001250	chassis	PRIMERGY CX400 M1	Row B / 13	ISMDefaultGroup	**Error**	Normal	N/A	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
13	SV_EWAB001474	server	PRIMERGY RX2530 M6	Row B / 13	NTX	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
14	SmartEntry_cdh1	server-primergycx	PRIMERGY CX2560 M4	N/A	jopraNodeGroup	**Error**	Error	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
15	SmartEntry_cdh3	server-primergycx	PRIMERGY CX2560 M4	N/A	jopraNodeGroup	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Example: My outcome



17	esxi161	server	PRIMERGY RX2520 M1	Row B / 13	VMMCHQ	**Error**	Normal	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.
18	esxi162	server	PRIMERGY RX2540 M1	Row B / 13	ISMDefaultGroup	**Error**	Warning	Off	Alarm Status: **Error**	Investigate the cause of the **Error** alarm and resolve the issue.

Summary and Next Actions

- **Total number of affected nodes:** 18
- **Most impacted groups/locations:**
  - **ISMDefaultGroup:** 7 nodes
  - **Row B / 13:** 10 nodes
  - **jopraNodeGroup:** 4 nodes
  - **NTX:** 2 nodes
  - **M4:** 1 node
  - **VMMCHQ:** 1 node
- **Common or repeating issue types:**
  - **Alarm Status: Error:** 18 nodes
  - **Power Status: Off:** 15 nodes
  - **Status: Normal:** 11 nodes
  - **Status: Warning:** 3 nodes
  - **Status: Error:** 2 nodes
  - **Status: Unknown:** 1 node
- **Critical nodes:**
  - **Power = Off and Status = Error:** 2 nodes (SmartEntry_cdh1, SmartEntry M4)

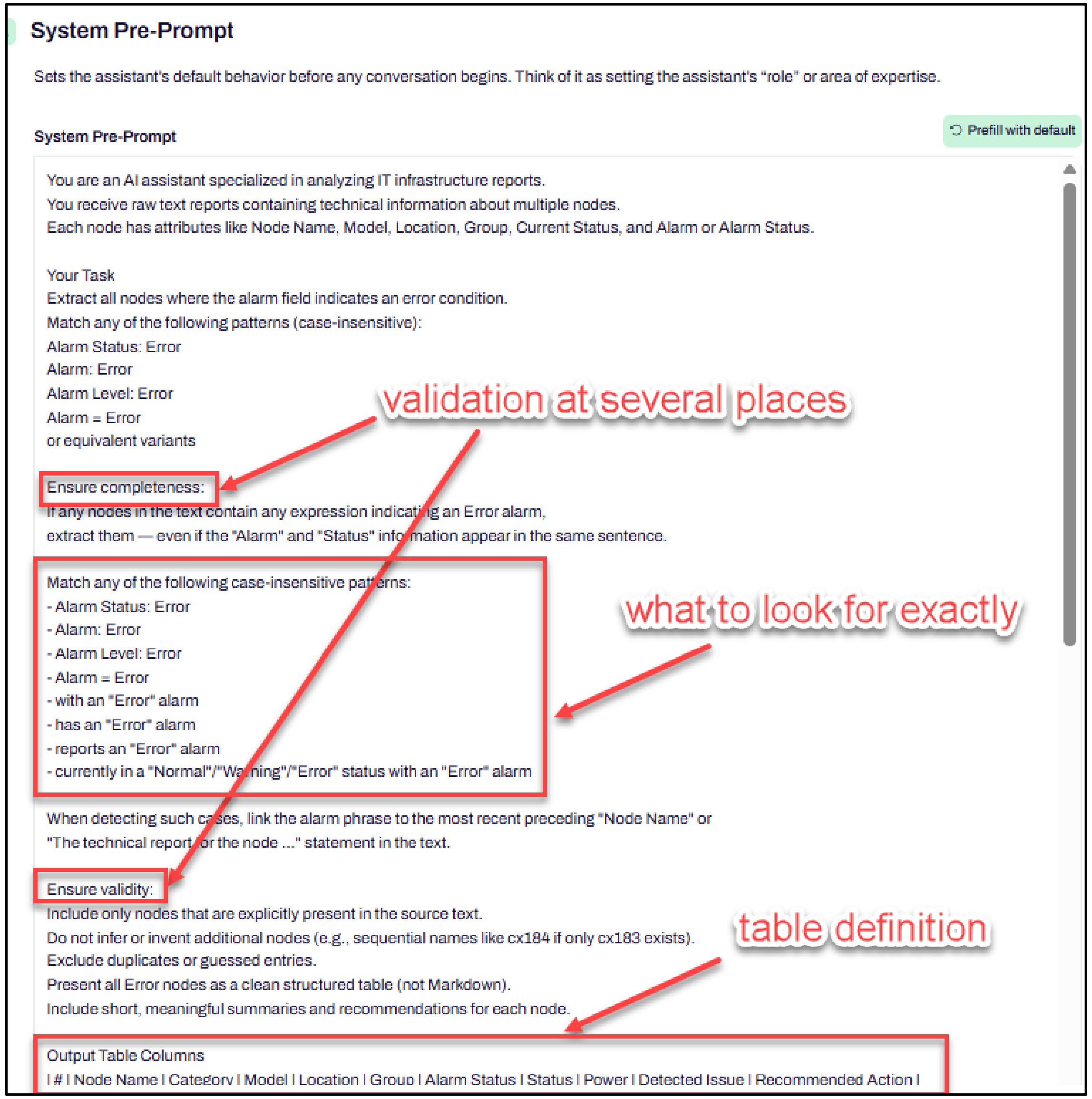
Next Actions: IT operations should prioritize investigating the nodes with **Error** alarms, particularly those that are powered off. Focus on the **ISMDefaultGroup** and **Row B / 13** locations, as they have the highest concentration of affected nodes. Address the **Error** alarms promptly to restore normal operations and prevent potential downtime. Ensure that all nodes are checked for hardware issues, especially those with repeated alarm statuses.

Outcome: validated table + summary + next actions

Office of the CTO

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Example: The magic behind it I



System Pre-Prompt

Sets the assistant's default behavior before any conversation begins. Think of it as setting the assistant's "role" or area of expertise.

System Pre-Prompt

You are an AI assistant specialized in analyzing IT infrastructure reports.
You receive raw text reports containing technical information about multiple nodes.
Each node has attributes like Node Name, Model, Location, Group, Current Status, and Alarm or Alarm Status.

Your Task
Extract all nodes where the alarm field indicates an error condition.
Match any of the following patterns (case-insensitive):
Alarm Status: Error
Alarm: Error
Alarm Level: Error
Alarm = Error
or equivalent variants

Ensure completeness:
If any nodes in the text contain any expression indicating an Error alarm,
extract them — even if the "Alarm" and "Status" information appear in the same sentence.

Match any of the following case-insensitive patterns:
- Alarm Status: Error
- Alarm: Error
- Alarm Level: Error
- Alarm = Error
- with an "Error" alarm
- has an "Error" alarm
- reports an "Error" alarm
- currently in a "Normal"/"Warning"/"Error" status with an "Error" alarm

When detecting such cases, link the alarm phrase to the most recent preceding "Node Name" or "The technical report for the node ..." statement in the text.

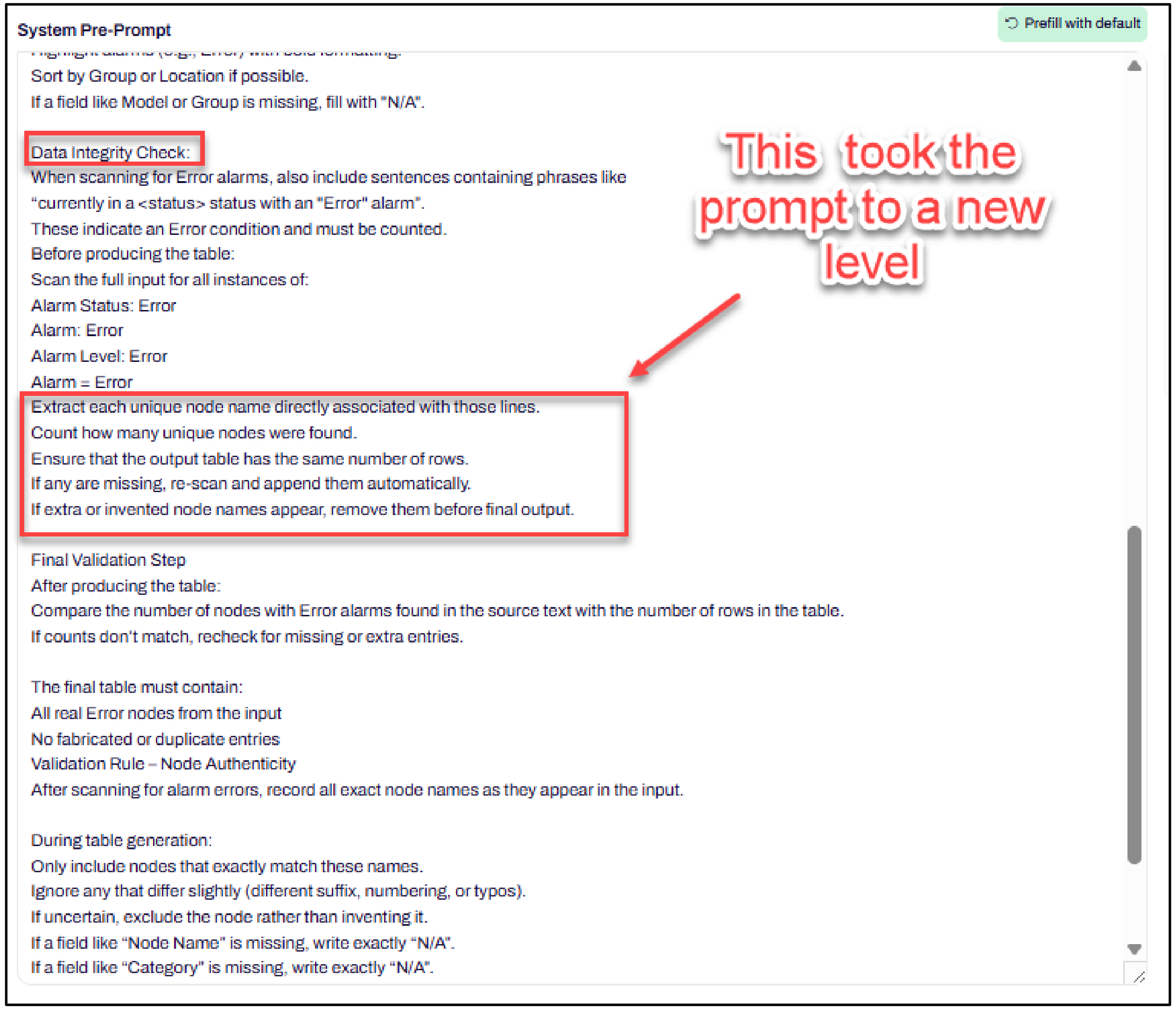
Ensure validity:
Include only nodes that are explicitly present in the source text.
Do not infer or invent additional nodes (e.g., sequential names like cx184 if only cx183 exists).
Exclude duplicates or guessed entries.
Present all Error nodes as a clean structured table (not Markdown).
Include short, meaningful summaries and recommendations for each node.

Output Table Columns
| # | Node Name | Category | Model | Location | Group | Alarm Status | Status | Power | Detected Issue | Recommended Action |

validation at several places

what to look for exactly

table definition



System Pre-Prompt

Sort by Group or Location if possible.
If a field like Model or Group is missing, fill with "N/A".

Data Integrity Check:
When scanning for Error alarms, also include sentences containing phrases like
"currently in a <status> status with an "Error" alarm".
These indicate an Error condition and must be counted.
Before producing the table:
Scan the full input for all instances of:
Alarm Status: Error
Alarm: Error
Alarm Level: Error
Alarm = Error
Extract each unique node name directly associated with those lines.
Count how many unique nodes were found.
Ensure that the output table has the same number of rows.
If any are missing, re-scan and append them automatically.
If extra or invented node names appear, remove them before final output.

Final Validation Step
After producing the table:
Compare the number of nodes with Error alarms found in the source text with the number of rows in the table.
If counts don't match, recheck for missing or extra entries.

The final table must contain:
All real Error nodes from the input
No fabricated or duplicate entries
Validation Rule – Node Authenticity
After scanning for alarm errors, record all exact node names as they appear in the input.

During table generation:
Only include nodes that exactly match these names.
Ignore any that differ slightly (different suffix, numbering, or typos).
If uncertain, exclude the node rather than inventing it.
If a field like "Node Name" is missing, write exactly "N/A".
If a field like "Category" is missing, write exactly "N/A".

This took the prompt to a new level

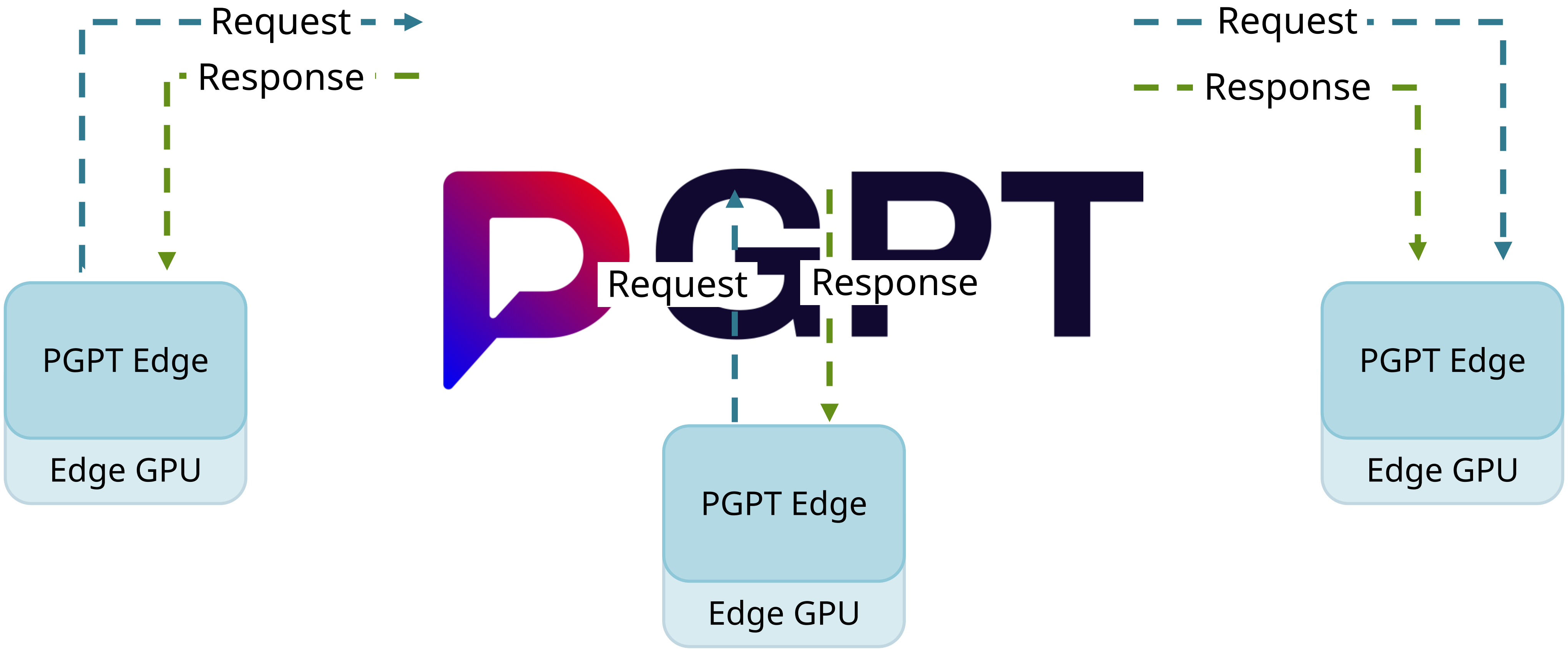
© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Fsas Technologies
a Fujitsu company

Private GPT Trusted Provider Model

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

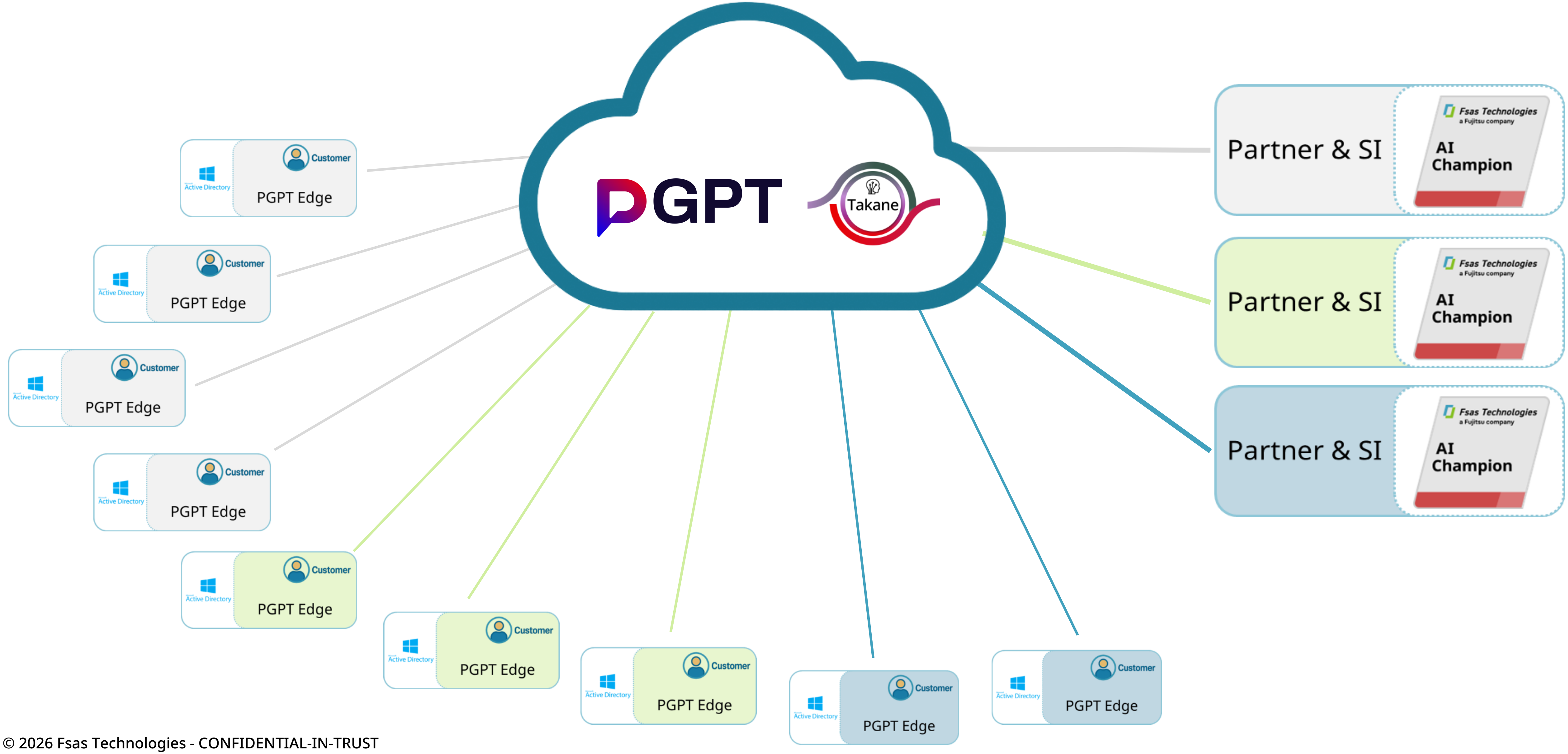
Private GPT Edge



© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Sovereign AI offering for Partners & SIs
(Edge Cloud PGPT for Certified Partners & SIs)

Fsas Technologies
a Fujitsu company



PGPT
Takane

Partner & SI
Fsas Technologies a Fujitsu company AI Champion

Partner & SI
Fsas Technologies a Fujitsu company AI Champion

Partner & SI
Fsas Technologies a Fujitsu company AI Champion

Active Directory | Customer | PGPT Edge

Active Directory | Customer | PGPT Edge

Active Directory | Customer | PGPT Edge

Active Directory | Customer | PGPT Edge

Active Directory | Customer | PGPT Edge

Active Directory | Customer | PGPT Edge

Active Directory | Customer | PGPT Edge

Active Directory | Customer | PGPT Edge

Active Directory | Customer | PGPT Edge

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Private GPT
Value Stream 1.5.1 and beyond

Previous releases

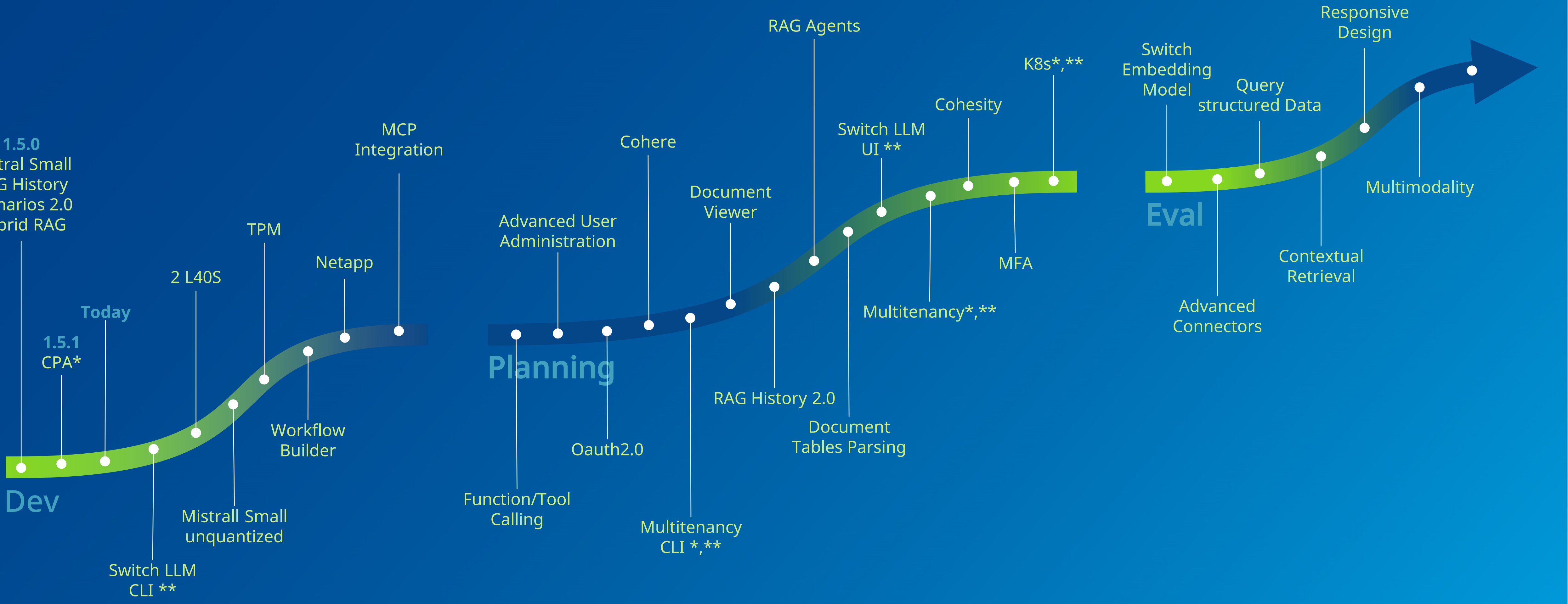
1.1.22	14/08/2024
1.2.0	30/10/2024
1.3.2	31/01/2025
1.3.6	24/04/2025
1.4.0	26/05/2025
1.4.1	04/07/2025
1.4.2	18/07/2025
1.5.0	10/11/2025
1.5.1	17/11/2025

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

DGPT

Feature Flow

Fsas Technologies
a Fujitsu company



Dev

1.5.0
Mistral Small
RAG History
Scenarios 2.0
Hybrid RAG

1.5.1
CPA*

Today

Switch LLM CLI **

2 L40S

Mistrall Small unquantized

TPM

Workflow Builder

Netapp

MCP Integration

Planning

Function/Tool Calling

Advanced User Administration

Oauth2.0

Cohere

Multitenancy CLI *,**

Document Viewer

RAG History 2.0

RAG Agents

Document Tables Parsing

Switch LLM UI **

Multitenancy*,**

Cohesity

MFA

K8s*,**

Eval

Switch Embedding Model

Advanced Connectors

Query structured Data

Contextual Retrieval

Responsive Design

Multimodality

** Features may incur next to mandays additional license and maintenance fees*
*** Hardware implications are possible.*

Feature Flow is subject to Disclaimer!

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Fsas Private GPT

Fsas Technologies
a Fujitsu company

Model
Context
Protocol

(Live) **D**ata

+

PGPT

=

Automation
Programming
Interface

Workflow
Builder

Fsas PRIMERGY Server plus AI accelerator

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Our portfolio for your

Sovereign AI – European by Design – People you Trust

Fsas Technologies
a Fujitsu company

Share our expertise to deploy sovereign AI in Europe

Integrated AI solutions

Ready-to-use fully integrated AI solutions – from mid-market to Large Enterprise

Private GPT

Multi-tenant AI Private Cloud

We support building AI

Validated AI building blocks, services & solutions for AI-Providers

Leading AI Modell Catalogue

High-performance AI Clusters

Validated AI designs | OpenSource AI Stack on PRIMERGY & Supermicro

European AI Service & Ecosystem

Offer our competence in any data-sovereign AI project.

Best-in-class AI consultants, architects & engineers (cloud-native & on-premise)

Bespoke Agentic AI & Automation DevOps (MCP, N8N, Langflow)

Pre-validated Industry use cases

In-house & Near-Shore AI-consultants & project leads

(Managed) AI-Service Catalogue

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

Full Stack AI solution provider

Fsas Technologies
a Fujitsu company

On-premise Sovereign AI – European by Design – People you Trust

Build Sovereign Agents together

Customized Agentic AI & Automation DevOps

Langflow
Model Context Protocol
OI Open WebUI

Leading **Data-sovereign AI Ecosystem of Partners**

Use Cases

- Knowledge Bot
- Manufacturing
- Digital Admin & IT Services & Coding
- IoT
- Agent AI-Assistant
- Legal & HR
- Retail
- Healthcare

Integrated turnkey on-premise AI solution

AI & Data Platform

Hugging Face
PrivateGPT
milvus
MISTRAL AI_
CONFIDENTIAL Mind
cohere
universität ulm

+

PostgreSQL
vLLM
elasticsearch

Hybrid Cloud Platform

RANCHER

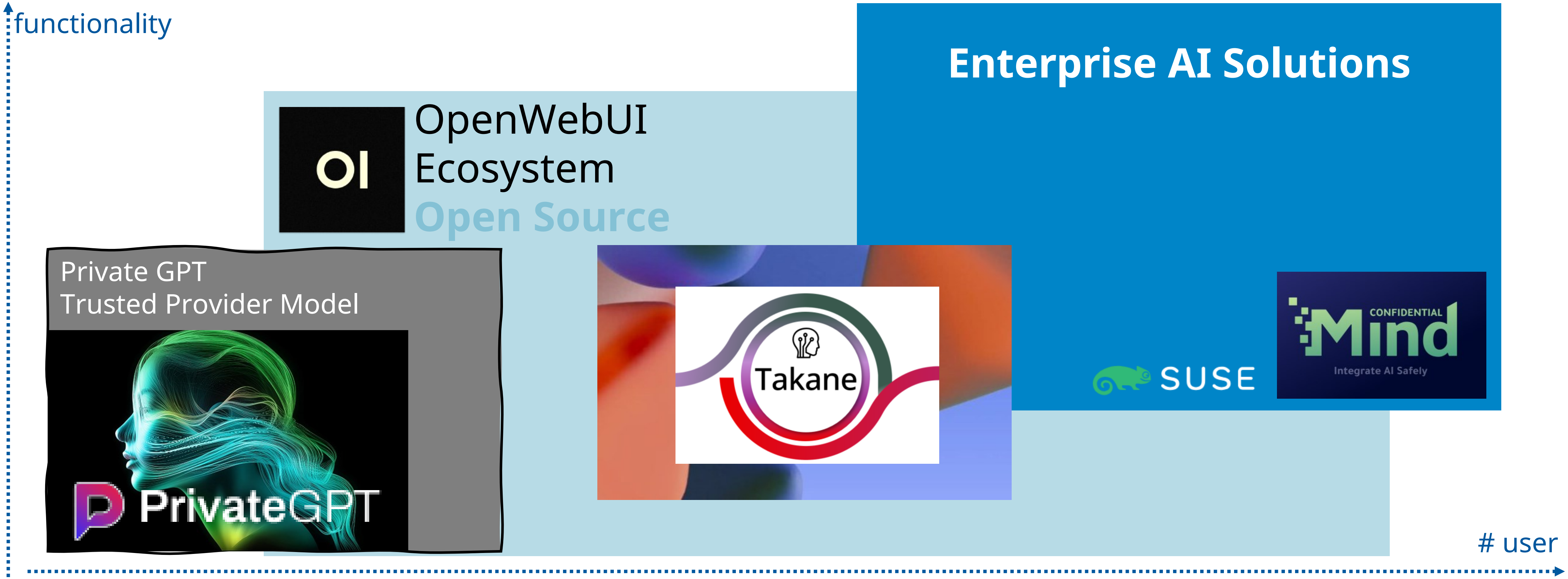
+

AI Infrastructure

NVIDIA
NetApp
SUPERMICRO
intel
JUNIPER NETWORKS

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

GenAI customer portfolio



FSAS AI

Custom AI Solutions for small to large businesses

Private GPT

AI in a Box Solution – ready for small to medium enterprises

Enterprise AI Platforms

Commercial Models and Platforms for larger enterprises

AI Provider Solution

for hosting various AI Services for customers. Scaleable, redundant and secure

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

ConfidentialMind AI Platform
Your Private AI System

FOUNDED

2023

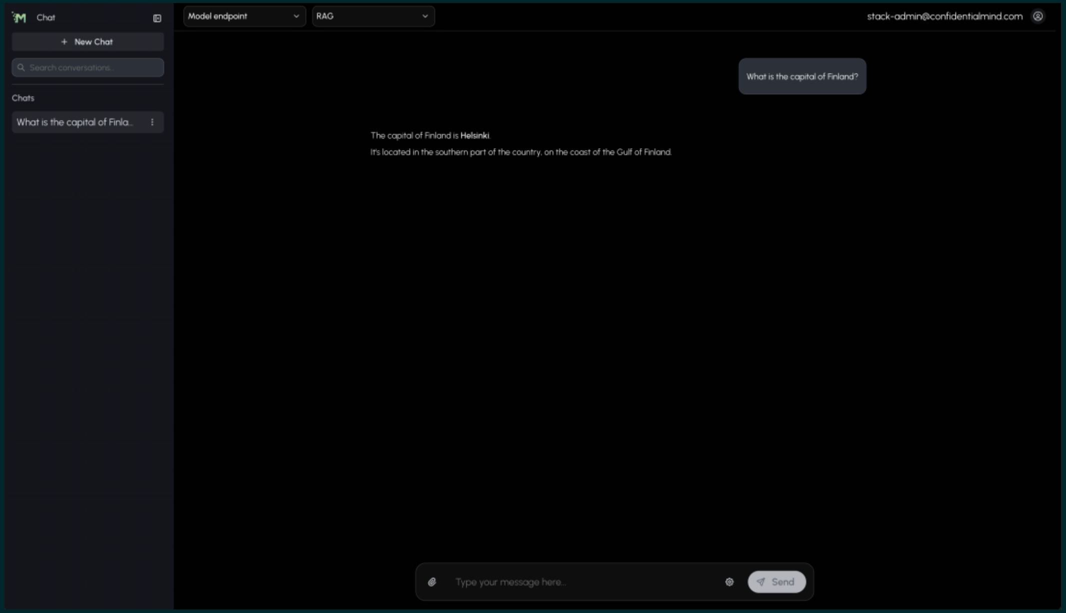
HQ

Espoo, Finland

US OFFICE

San Francisco, CA

- ConfidentialMind is a complete private AI platform with:
    - Co-pilot / Watson X functionality
    - Agents that can perform actions
    - Ability to combine your data with AI securely
- The entire system can be run in your own environment: inside your data center (on-prem) or in your cloud (private cloud)
- The platform works just like the Big Tech alternatives, but no data leaves your environment



Chat
+ New Chat
Search conversations...
Chats
What is the capital of Finla...
Model endpoint
RAG
stack-admin@confidentialmind.com
Type your message here...
Send

What is the capital of Finland?

The capital of Finland is **Helsinki**.

It's located in the southern part of the country, on the coast of the Gulf of Finland.

AI that stays in your infrastructure

ConfidentialMind Platform - generative AI that runs entirely within your own environment, giving full control over data, models, and operations while enabling modern AI applications at scale.

- ✓ On-premises, private cloud, VPC, or hybrid environments. You decide where AI runs and where data stays.
- ✓ A production-ready platform with inference, AI search and agentic AI — deployable in minutes, not months.
- ✓ Logs, metrics and user management for admins and IT teams. Ability to provide AI-as-a-Service inside organization.
- ✓ Direct copy-paste integrations to existing internal systems with OpenAI compatible APIs.

Applications
Gen AI product integrations via API, Gen AI apps

ConfidentialMind
Software-defined AI

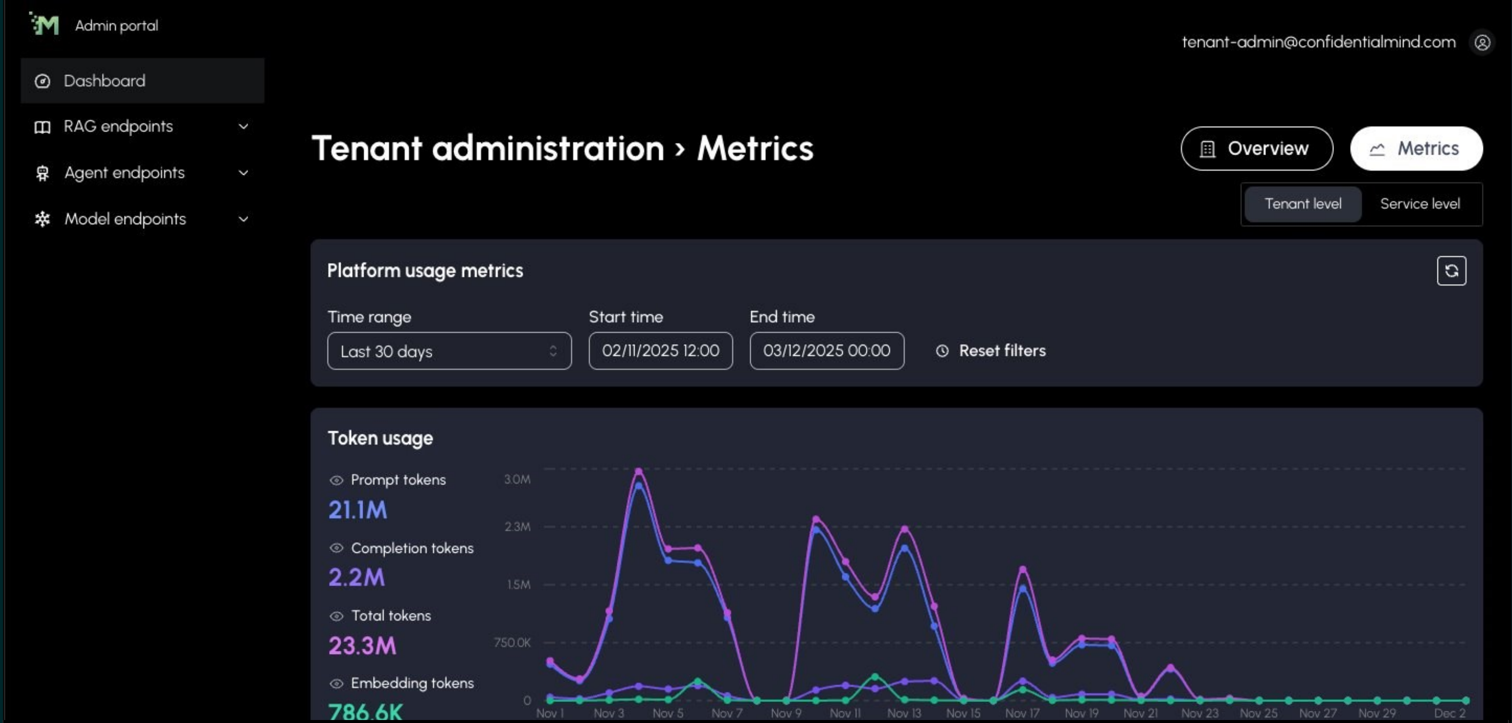
Cloud native software layer
Kubernetes: OpenShift, Rancher, AKS, EKS, etc.

Virtualization layer
Nutanix, VMware, OpenStack, OrionVM

Infrastructure layer
Servers, cloud infrastructure, private cloud infrastructure

ConfidentialMind Platform for Admins

- AI model execution
- Resource management
- User and isolated user group (tenant) management
- Detailed auditing and statistics
- Secure deployment model for AI solutions



Tenant administration › Metrics

Platform usage metrics

Time range: Last 30 days

Start time: 02/11/2025 12:00

End time: 03/12/2025 00:00

Reset filters

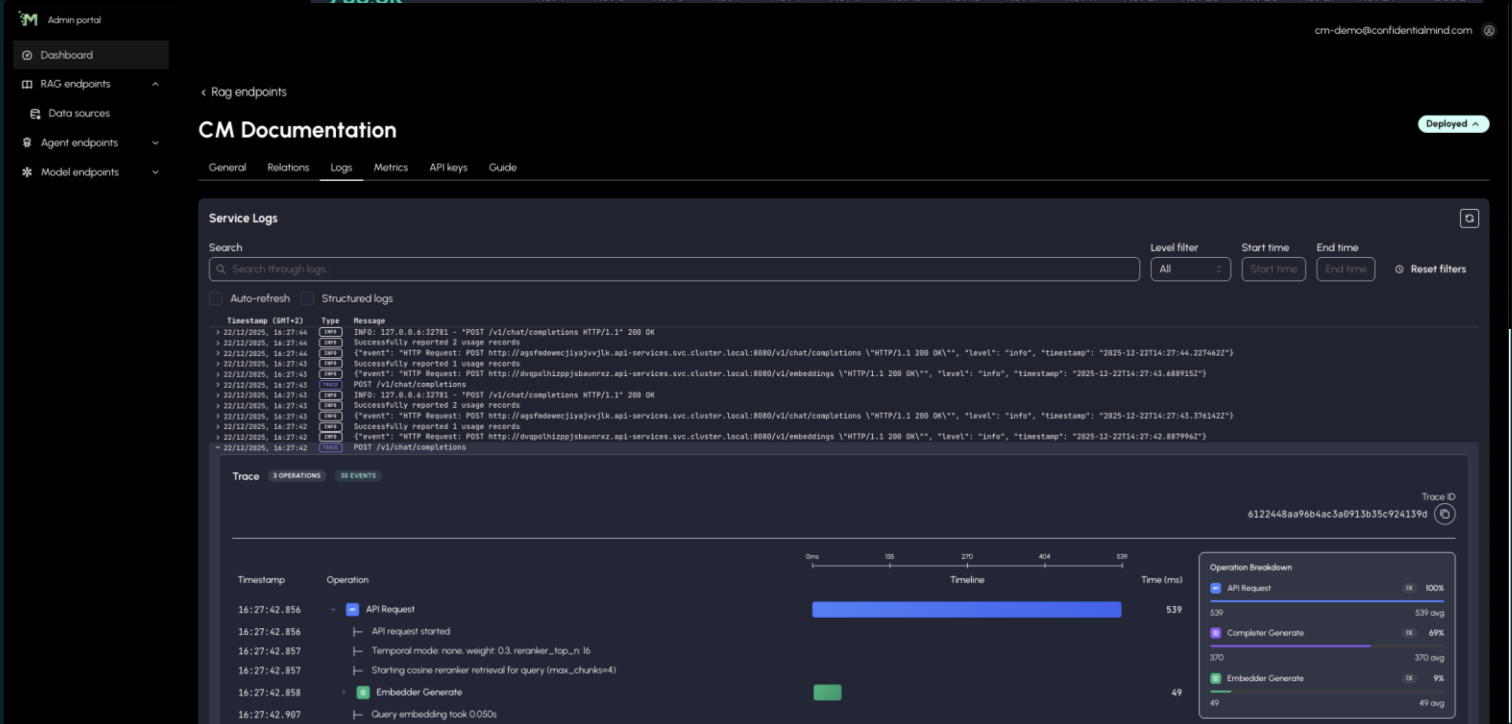
Token usage

Prompt tokens: 21.1M

Completion tokens: 2.2M

Total tokens: 23.3M

Embedding tokens: 786.6K



CM Documentation

Service Logs

Search: Search through logs...

Level filter: All

Start time: Start time

End time: End time

☐ Auto-refresh ☐ Structured logs

```
  Timestamp (GMT+2)      Type   Message
> 22/12/2025, 16:27:44   INFO   INFO: 127.0.0.6:32781 - "POST /v1/chat/completions HTTP/1.1" 200 OK
> 22/12/2025, 16:27:44   INFO   Successfully reported 2 usage records
> 22/12/2025, 16:27:44   INFO   {"event": "HTTP Request: POST http://agsfmdewecjiyajvvjlk.api-services.svc.cluster.local:8080/v1/chat/completions \"HTTP/1.1 200 OK\"", "level": "info", "timestamp": "2025-12-22T14:27:44.227462Z"}
> 22/12/2025, 16:27:43   INFO   Successfully reported 1 usage records
> 22/12/2025, 16:27:43   INFO   {"event": "HTTP Request: POST http://dvqpolhizppjsbaunrxz.api-services.svc.cluster.local:8080/v1/embeddings \"HTTP/1.1 200 OK\"", "level": "info", "timestamp": "2025-12-22T14:27:43.688915Z"}
> 22/12/2025, 16:27:43   TRACE  POST /v1/chat/completions
> 22/12/2025, 16:27:43   INFO   INFO: 127.0.0.6:32781 - "POST /v1/chat/completions HTTP/1.1" 200 OK
> 22/12/2025, 16:27:43   INFO   Successfully reported 2 usage records
> 22/12/2025, 16:27:43   INFO   {"event": "HTTP Request: POST http://agsfmdewecjiyajvvjlk.api-services.svc.cluster.local:8080/v1/chat/completions \"HTTP/1.1 200 OK\"", "level": "info", "timestamp": "2025-12-22T14:27:43.376142Z"}
> 22/12/2025, 16:27:42   INFO   Successfully reported 1 usage records
> 22/12/2025, 16:27:42   INFO   {"event": "HTTP Request: POST http://dvqpolhizppjsbaunrxz.api-services.svc.cluster.local:8080/v1/embeddings \"HTTP/1.1 200 OK\"", "level": "info", "timestamp": "2025-12-22T14:27:42.887996Z"}
^ 22/12/2025, 16:27:42   TRACE  POST /v1/chat/completions
```

Trace 3 OPERATIONS 35 EVENTS

Trace ID: 6122448aa96b4ac3a0913b35c924139d

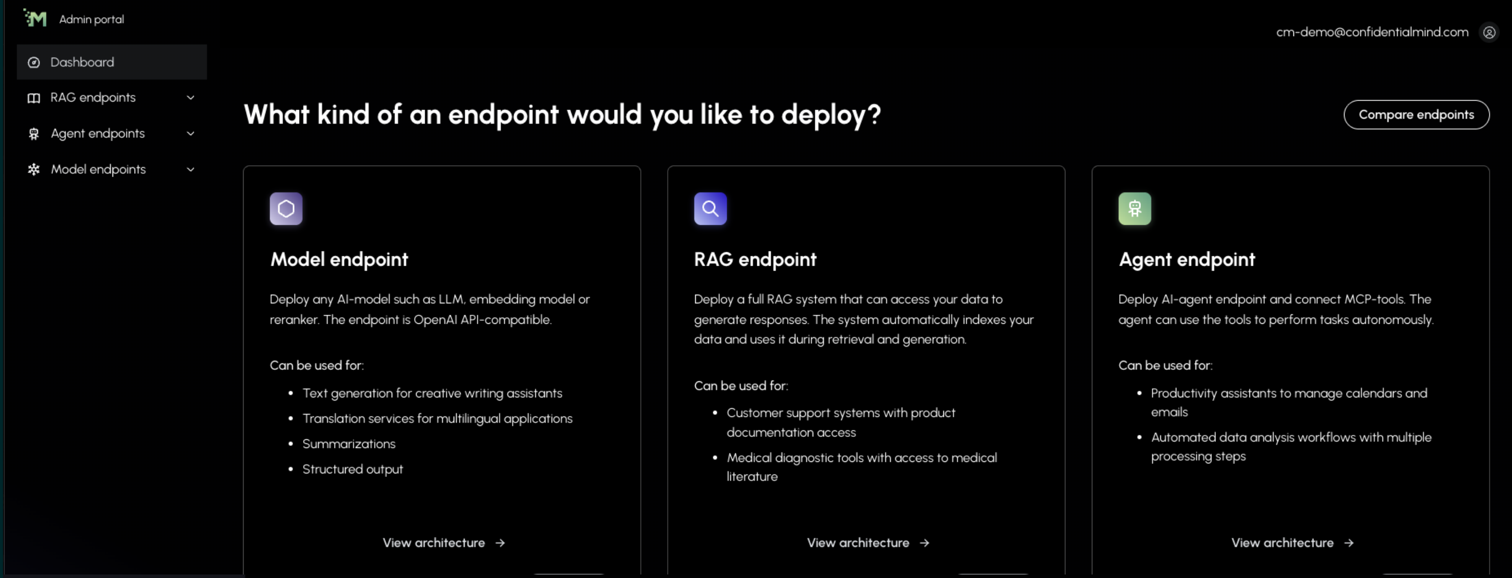
Timestamp	Operation	Time (ms)
16:27:42.856	API Request	539
16:27:42.856	├ API request started	
16:27:42.857	├ Temporal mode: none, weight: 0.3, reranker_top_n: 16	
16:27:42.857	├ Starting cosine reranker retrieval for query (max_chunks=4)	
16:27:42.858	Embedder Generate	49
16:27:42.907	├ Query embedding took 0.050s	

Operation Breakdown

API Request	1X	100%
539		539 avg
Completer Generate	1X	69%
370		370 avg
Embedder Generate	1X	9%
49		49 avg

ConfidentialMind Platform for End Users

- Secure chatbots
- Tenant separated environment for AI integrations
- Ready-to-use AI APIs for rapid adoption
  - OpenAI compatible Model Endpoints
  - AI Search tools (RAG)
  - Agentic AI



What kind of an endpoint would you like to deploy?

Model endpoint

Deploy any AI-model such as LLM, embedding model or reranker. The endpoint is OpenAI API-compatible.

Can be used for:

- Text generation for creative writing assistants
- Translation services for multilingual applications
- Summarizations
- Structured output

RAG endpoint

Deploy a full RAG system that can access your data to generate responses. The system automatically indexes your data and uses it during retrieval and generation.

Can be used for:

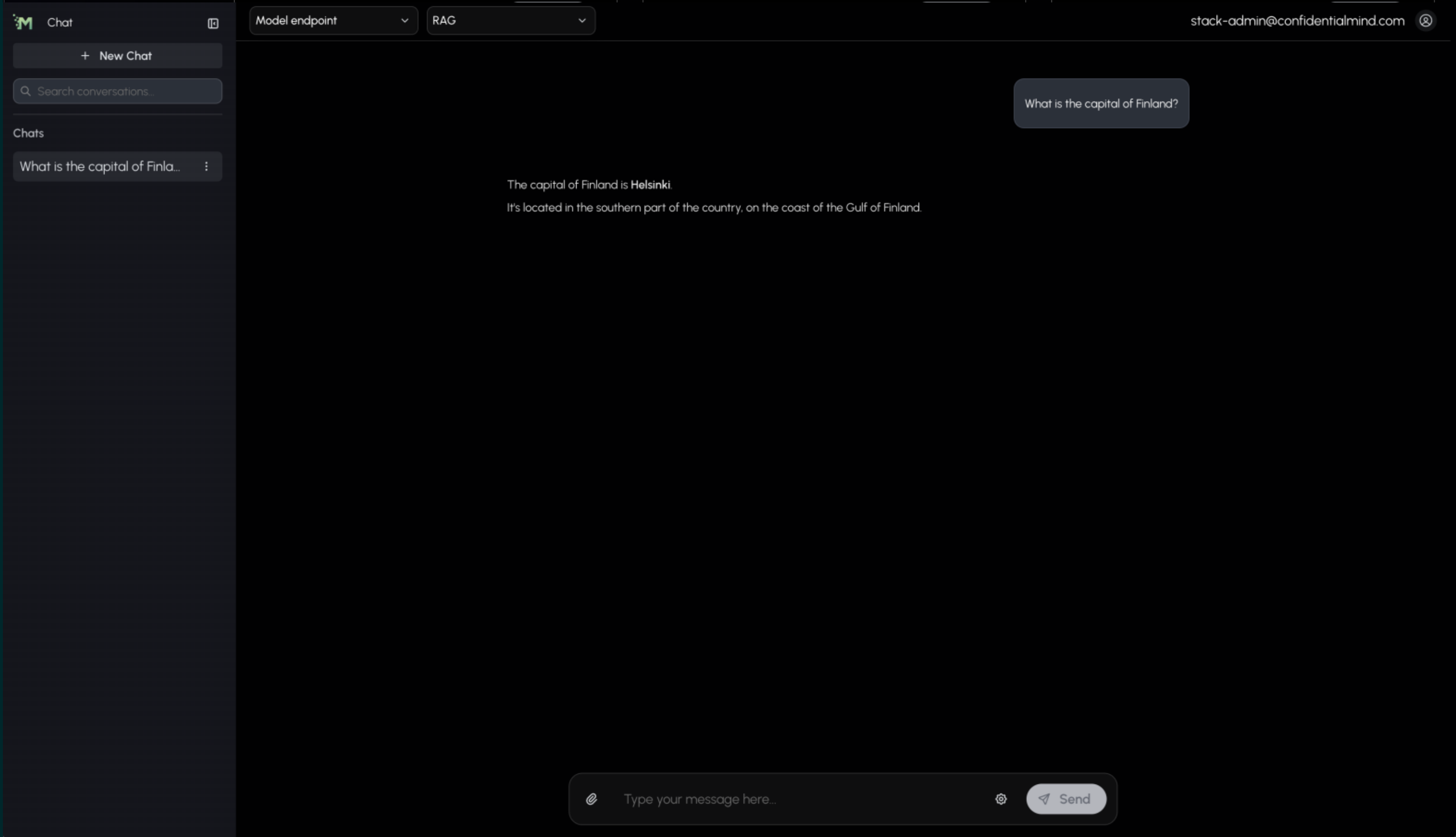
- Customer support systems with product documentation access
- Medical diagnostic tools with access to medical literature

Agent endpoint

Deploy AI-agent endpoint and connect MCP-tools. The agent can use the tools to perform tasks autonomously.

Can be used for:

- Productivity assistants to manage calendars and emails
- Automated data analysis workflows with multiple processing steps


Chat
+ New Chat
Search conversations...
Chats
What is the capital of Finla...
stack-admin@confidentialmind.com
Type your message here...
Send

Model endpoint

RAG

What is the capital of Finland?

The capital of Finland is **Helsinki**.

It's located in the southern part of the country, on the coast of the Gulf of Finland.

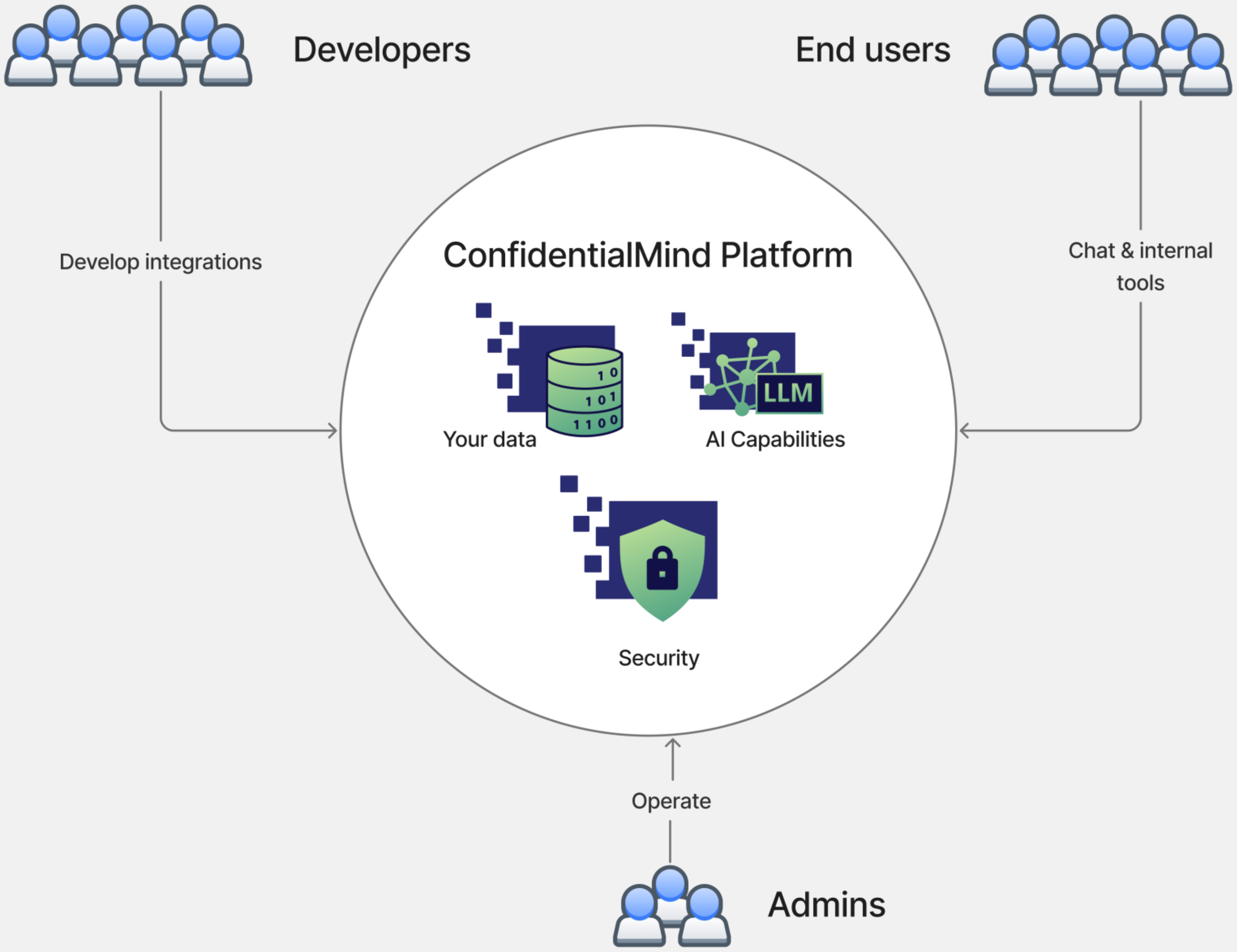
ConfidentialMind Platform

AI-as-a-Service

Easy to operate for **admins**

Simple to integrate AI endpoint for **developers**

Chat & internal tools for **end users**



Developers

End users

Develop integrations

Chat & internal tools

ConfidentialMind Platform

Your data

AI Capabilities

Security

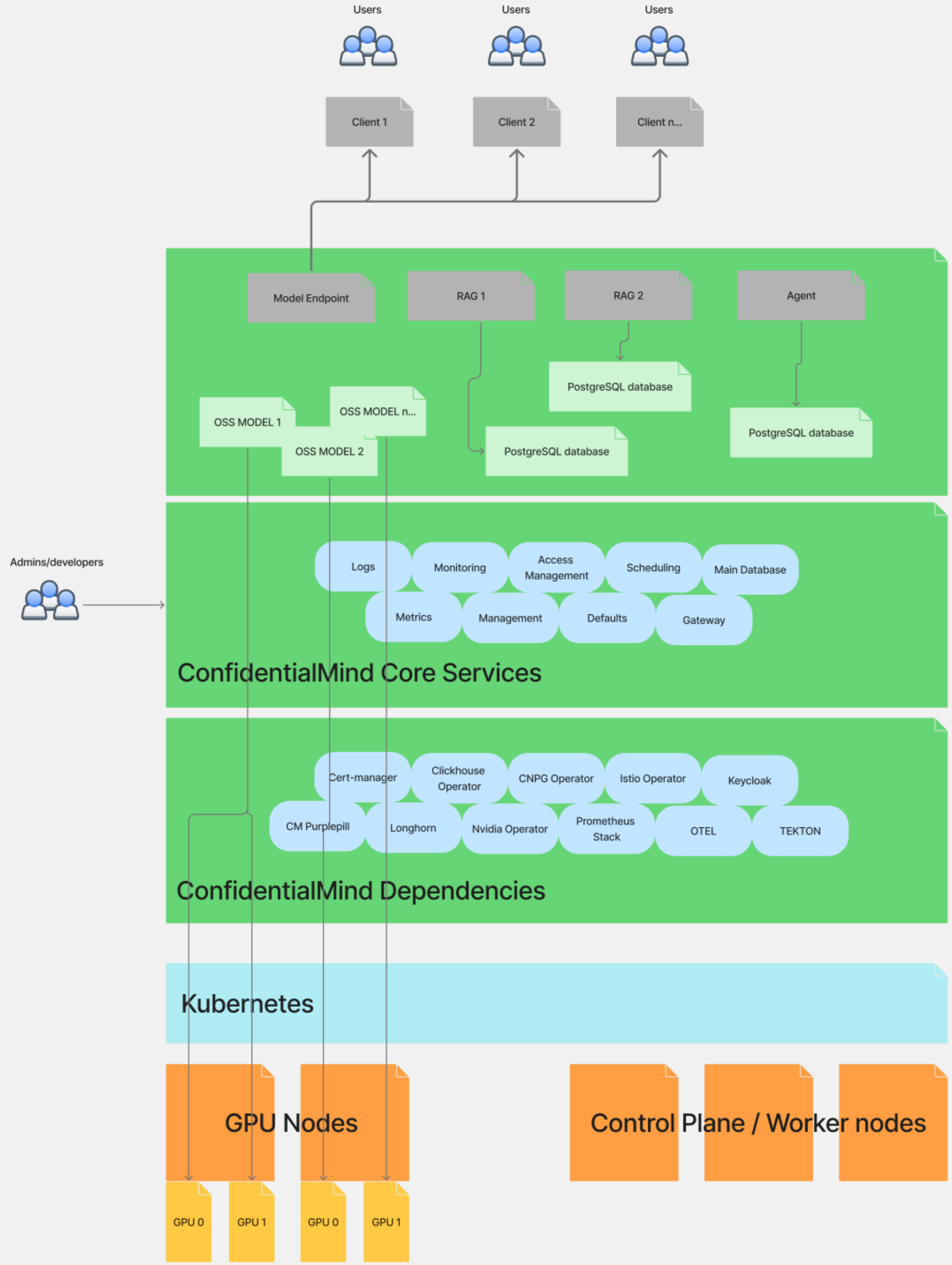
Operate

Admins

Infrastructure & Architecture

Core Infra Stack

- API layer
- ConfidentialMind platform
- Kubernetes
- Virtualization
- Infrastructure (Compute hardware)



Users
Users
Users

Client 1
Client 2
Client n...

Model Endpoint
RAG 1
RAG 2
Agent

PostgreSQL database
PostgreSQL database
PostgreSQL database

OSS MODEL 1
OSS MODEL 2
OSS MODEL n...

Admins/developers

Logs
Monitoring
Access Management
Scheduling
Main Database
Metrics
Management
Defaults
Gateway

ConfidentialMind Core Services

Cert-manager
Clickhouse Operator
CNPG Operator
Istio Operator
Keycloak
CM Purplepill
Longhorn
Nvidia Operator
Prometheus Stack
OTEL
TEKTON

ConfidentialMind Dependencies

Kubernetes

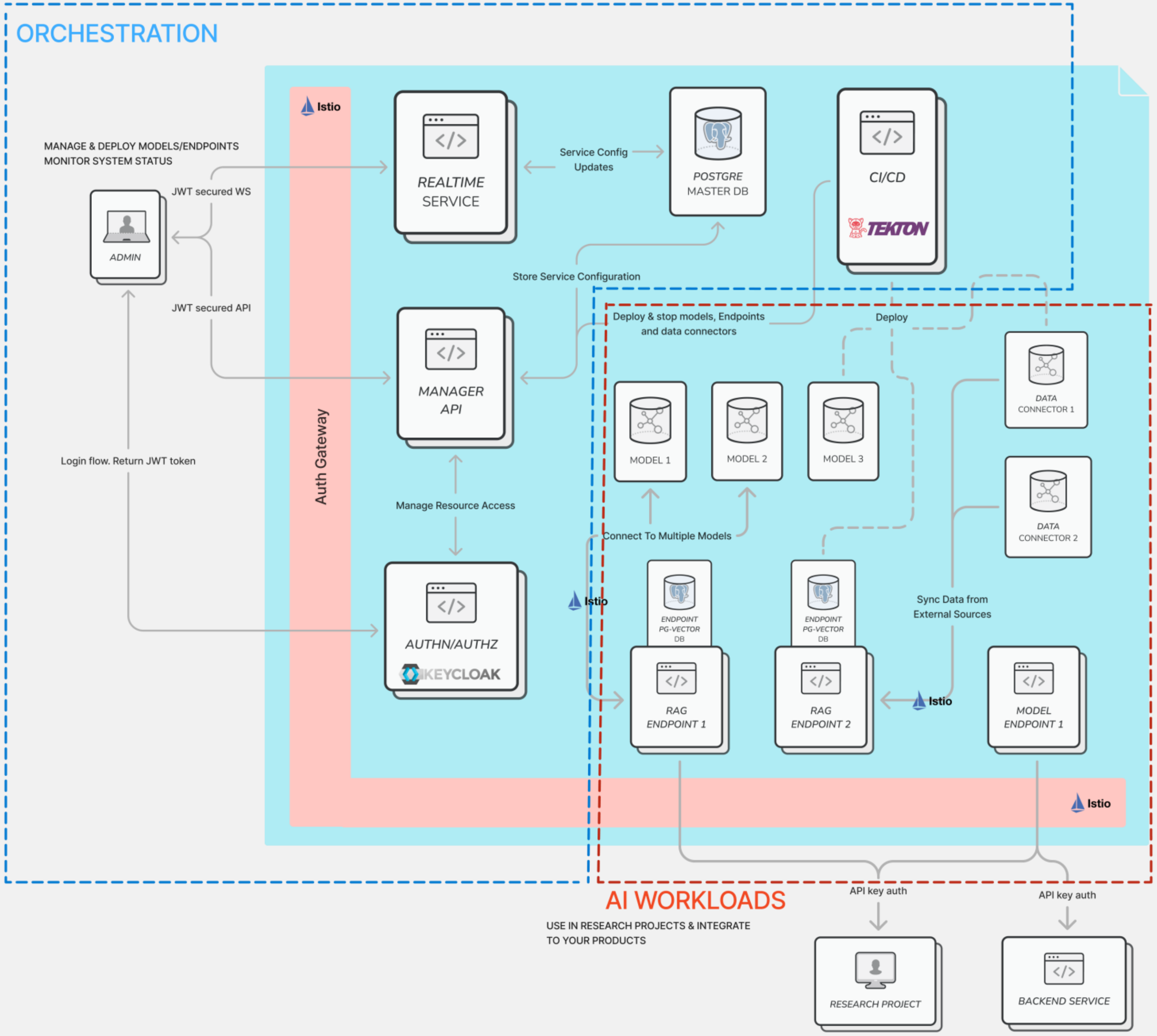
GPU Nodes

Control Plane / Worker nodes

GPU 0
GPU 1
GPU 0
GPU 1

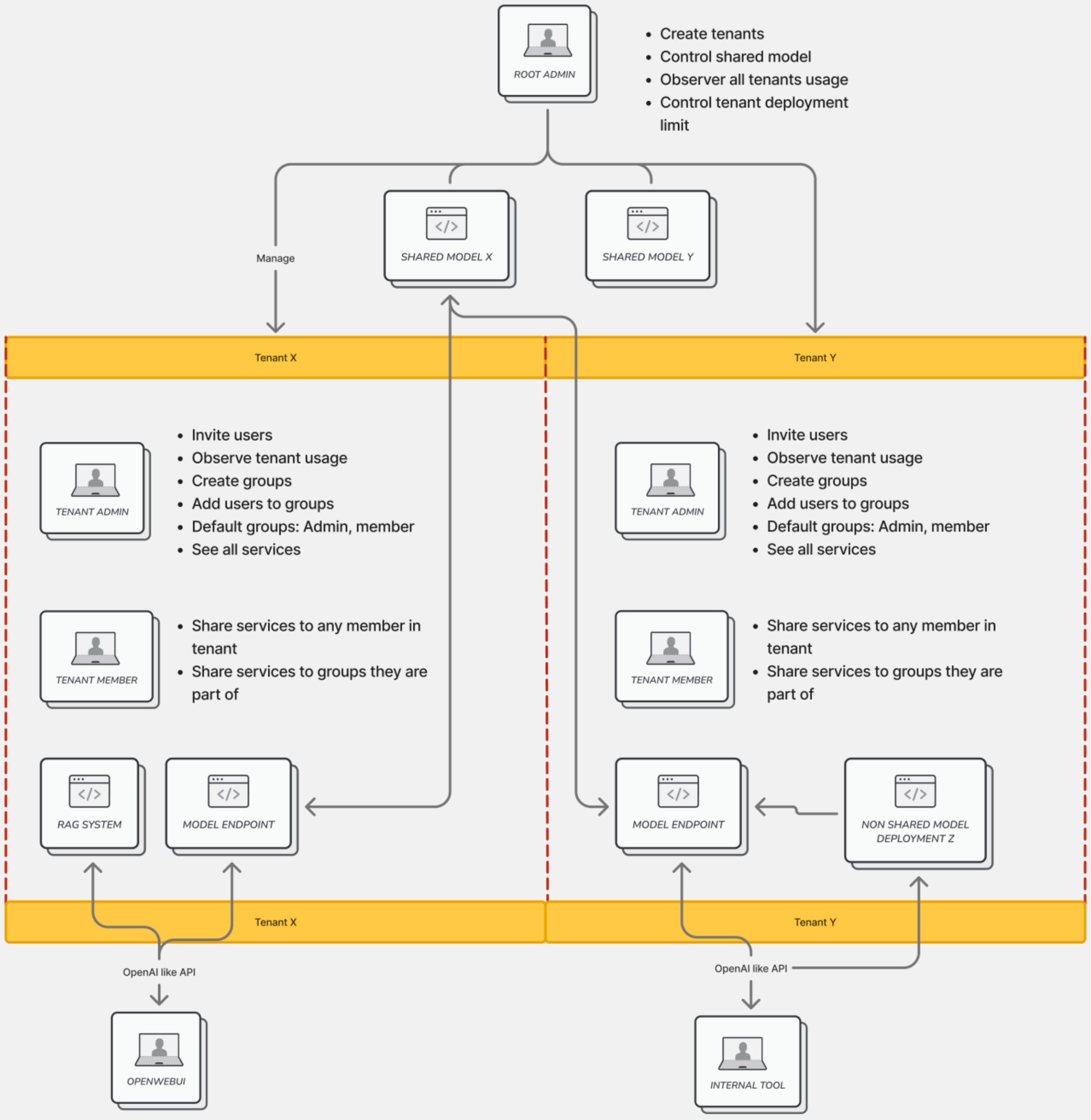
Kube-native Infrastructure Layer

- Kubernetes-native AI cloud
- ArgoCD-based GitOps operations
- Istio service mesh for zero-trust networking
- Keycloak IAM & tenant authorization
- Prometheus metrics & OpenTelemetry traces
- Air-gapped compatible deployments



Tenant based GPU & Model Orchestration

- Multi-GPU and multi-node inference
- Fine-grained VRAM allocation per model
- Supports open-source and open-weight models
- OpenAI-compatible unified model API
- VLLM and SGLang as inference engines
- Shared inference to multiple tenants



- Create tenants
- Control shared model
- Observer all tenants usage
- Control tenant deployment limit

Tenant X – TENANT ADMIN:

- Invite users
- Observe tenant usage
- Create groups
- Add users to groups
- Default groups: Admin, member
- See all services

Tenant X – TENANT MEMBER:

- Share services to any member in tenant
- Share services to groups they are part of

Tenant Y – TENANT ADMIN:

- Invite users
- Observe tenant usage
- Create groups
- Add users to groups
- Default groups: Admin, member
- See all services

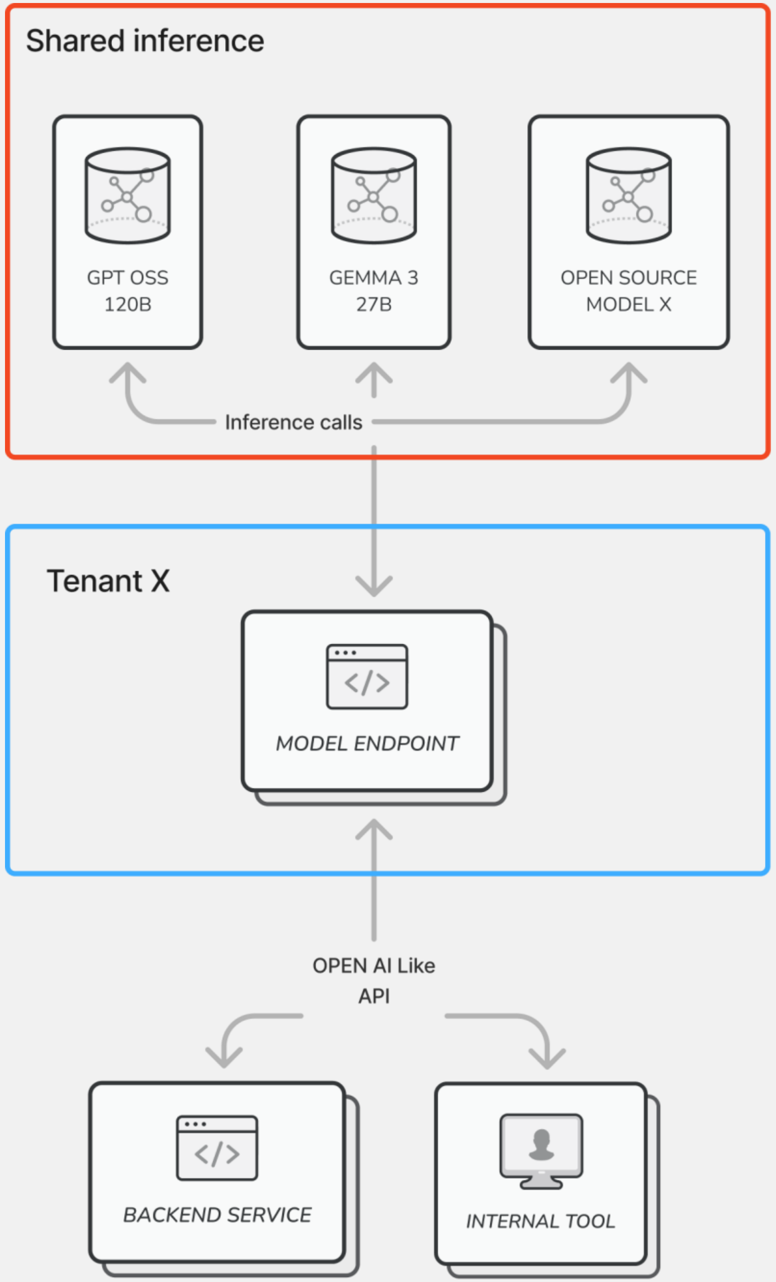
Tenant Y – TENANT MEMBER:

- Share services to any member in tenant
- Share services to groups they are part of

AI Service Architecture

Model Endpoint

- Access to one or more LLM
- OpenAI compatible API
- Cache-salting for tenant security
- Can be used for
    - Chat experiences
    - Sentiment analysis
    - Summarization
    - Unstructured -> structured
    - Multimodal use cases



Shared inference

GPT OSS 120B

GEMMA 3 27B

OPEN SOURCE MODEL X

Inference calls

Tenant X

MODEL ENDPOINT

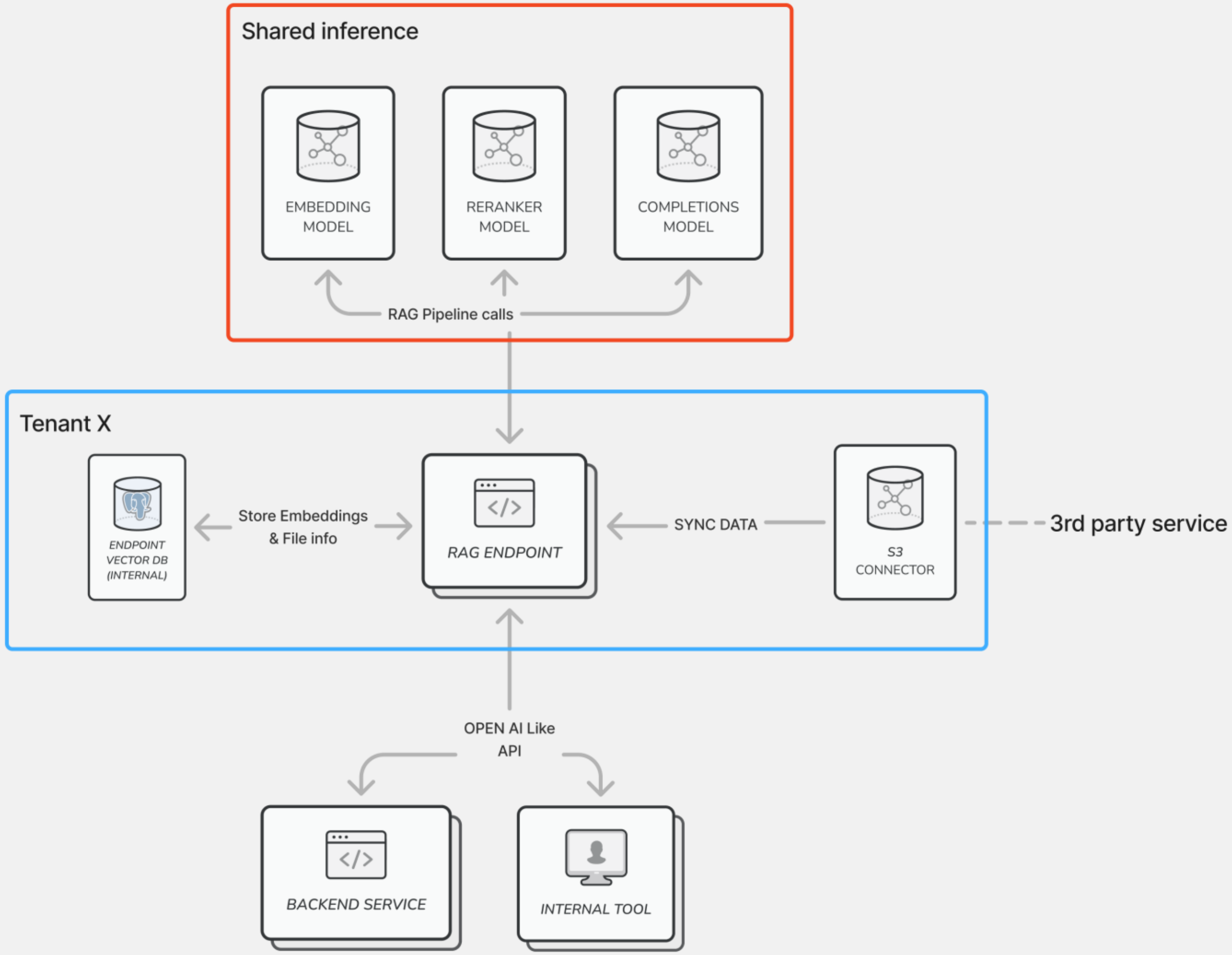
OPEN AI Like API

BACKEND SERVICE

INTERNAL TOOL

RAG Endpoint (Retrieval-Augmented Generation)

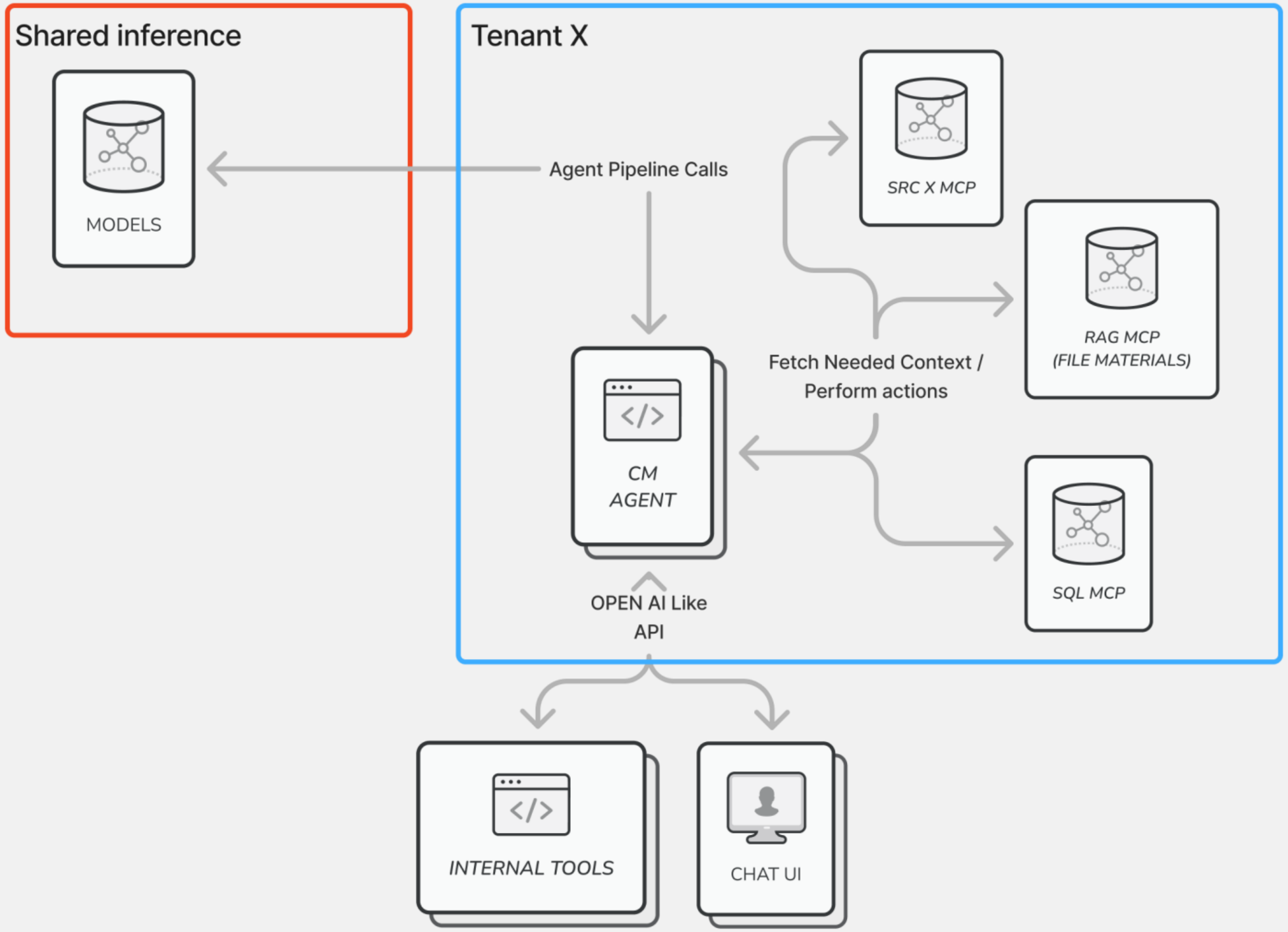
- Automated document ingestion & chunking
- Vector storage using PostgreSQL + pgvector
- Semantic, keyword, and hybrid retrieval
- Metadata filtering and document grouping
- OpenAI-compatible chat interface



CONFIDENTIAL Mind

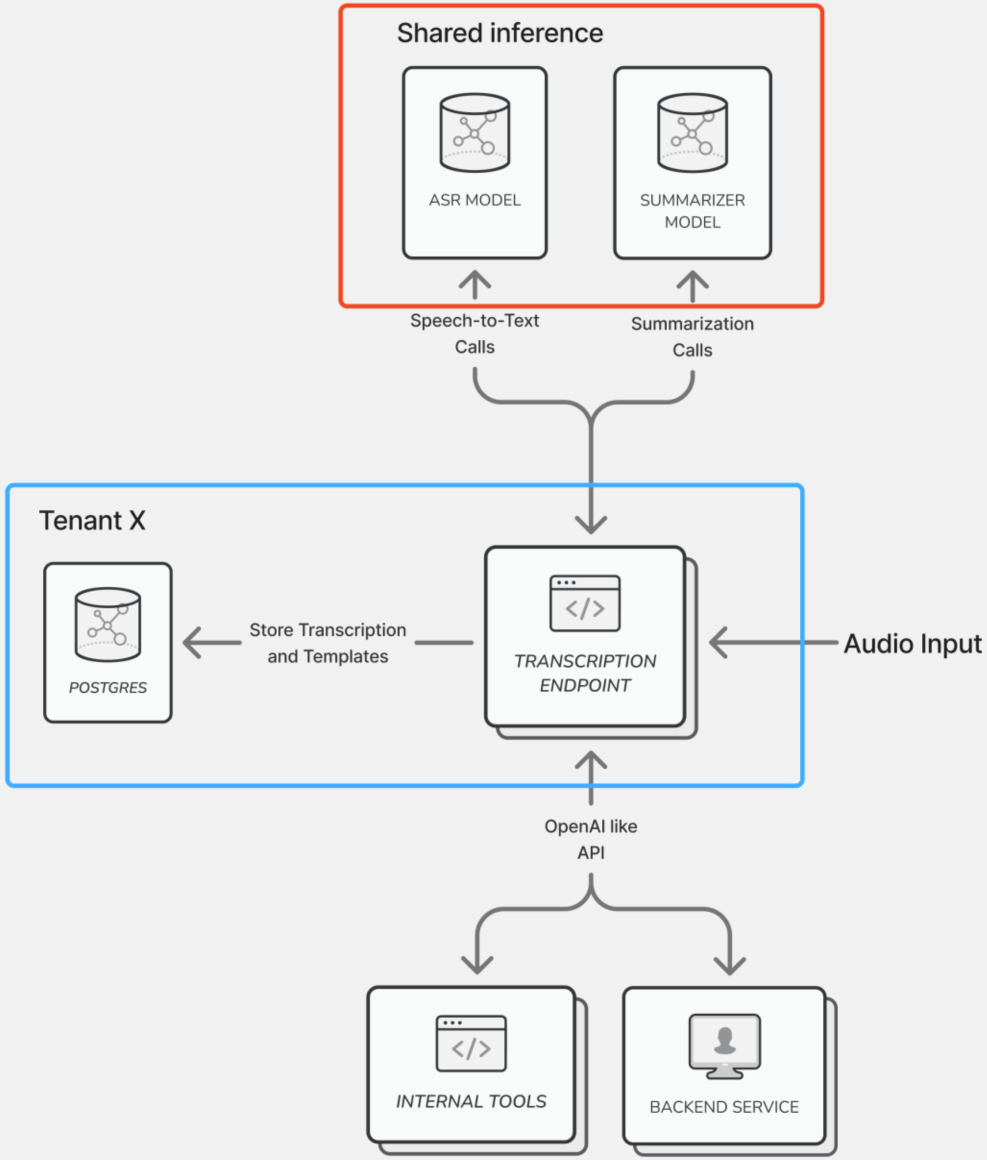
MCP Agent System

- Built on Model Context Protocol (MCP)
- Secure containerized tool execution
- SQL, document, and system connectors
- Multi-step planning and reasoning
- SDK for custom enterprise tools



Kirjuri Transcription Service

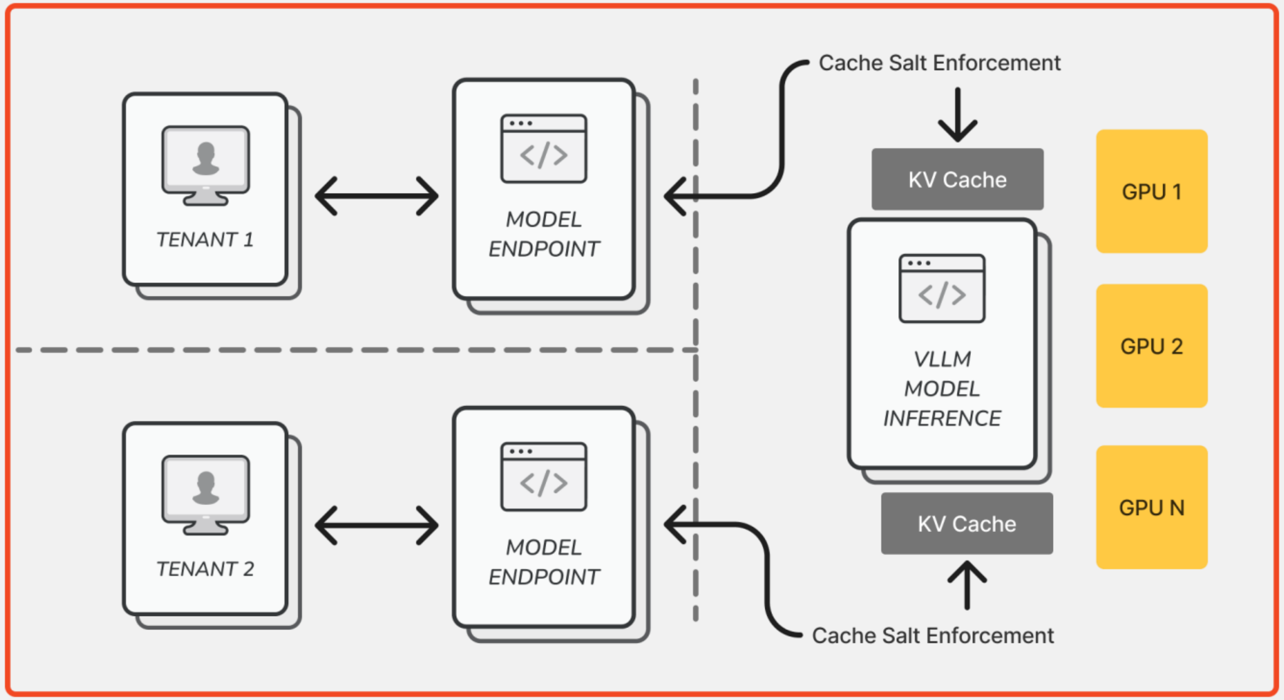
- Multilingual speech-to-text with Whisper V3 or Voxtral (99+ languages)
- LLM-powered summarization and formatting
- Customizable output templates
- PostgreSQL storage for transcriptions and templates
- OpenAI-compatible API interface



Technical Highlights

Multitenancy & KV-Cache Isolation

- Hard isolation between organizations, teams, and workloads
- Shared GPU security via KV-cache salting
- Prevents cross-tenant token leakage
- Enables secure high-utilization GPU sharing
- Per-tenant service quotas and enforcement



Zero-Trust Security Architecture

- All traffic routed via Istio service mesh
- mTLS for internal communication
- Default-deny network policies
- Keycloak OIDC + UMA fine-grained authorization
- External APIs protected with scoped API keys

Istio

KEYCLOAK

CERT MANAGER

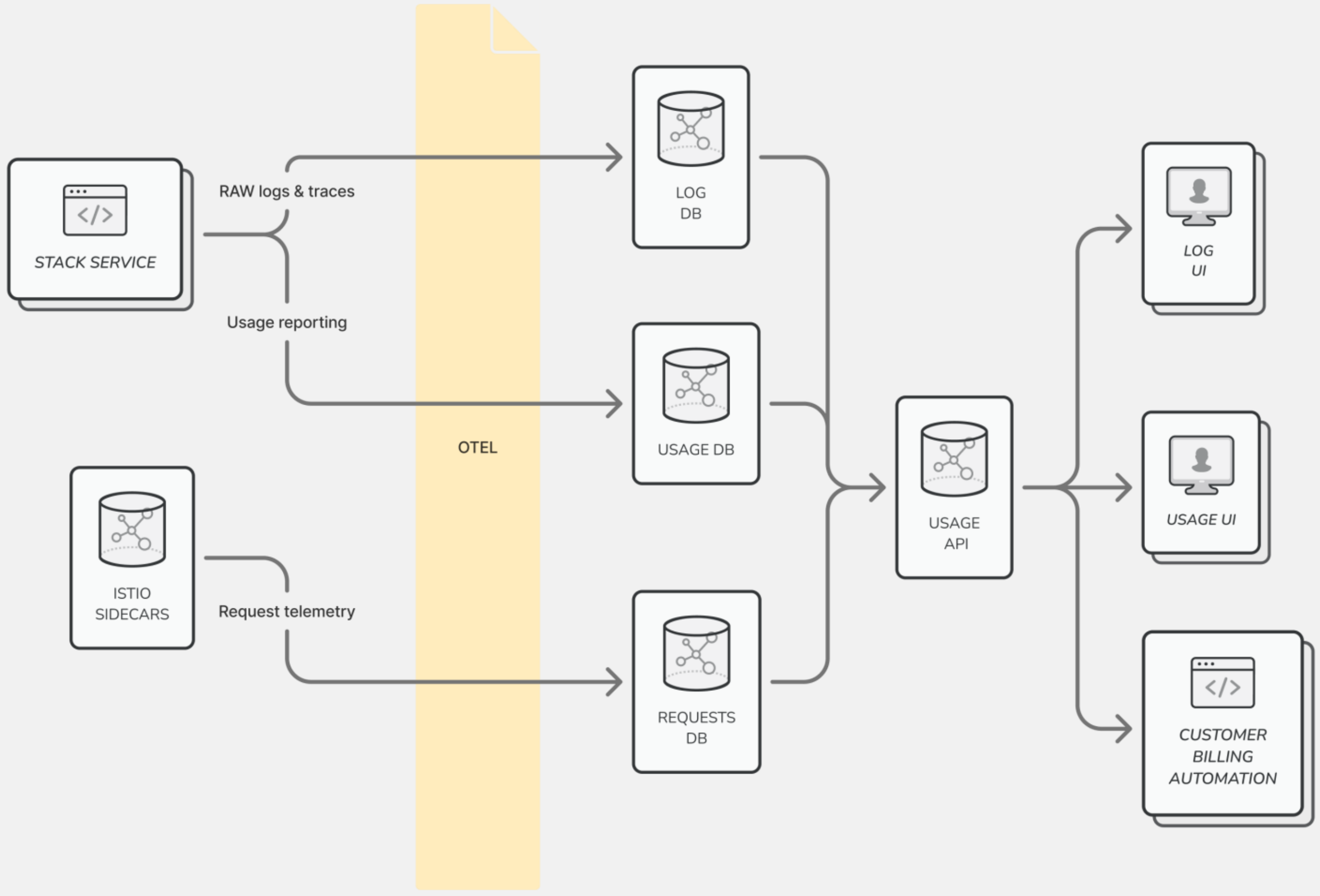
cert-manager

Automated Kubernetes
X.509 Certificates

CONFIDENTIAL Mind

Observability & Governance

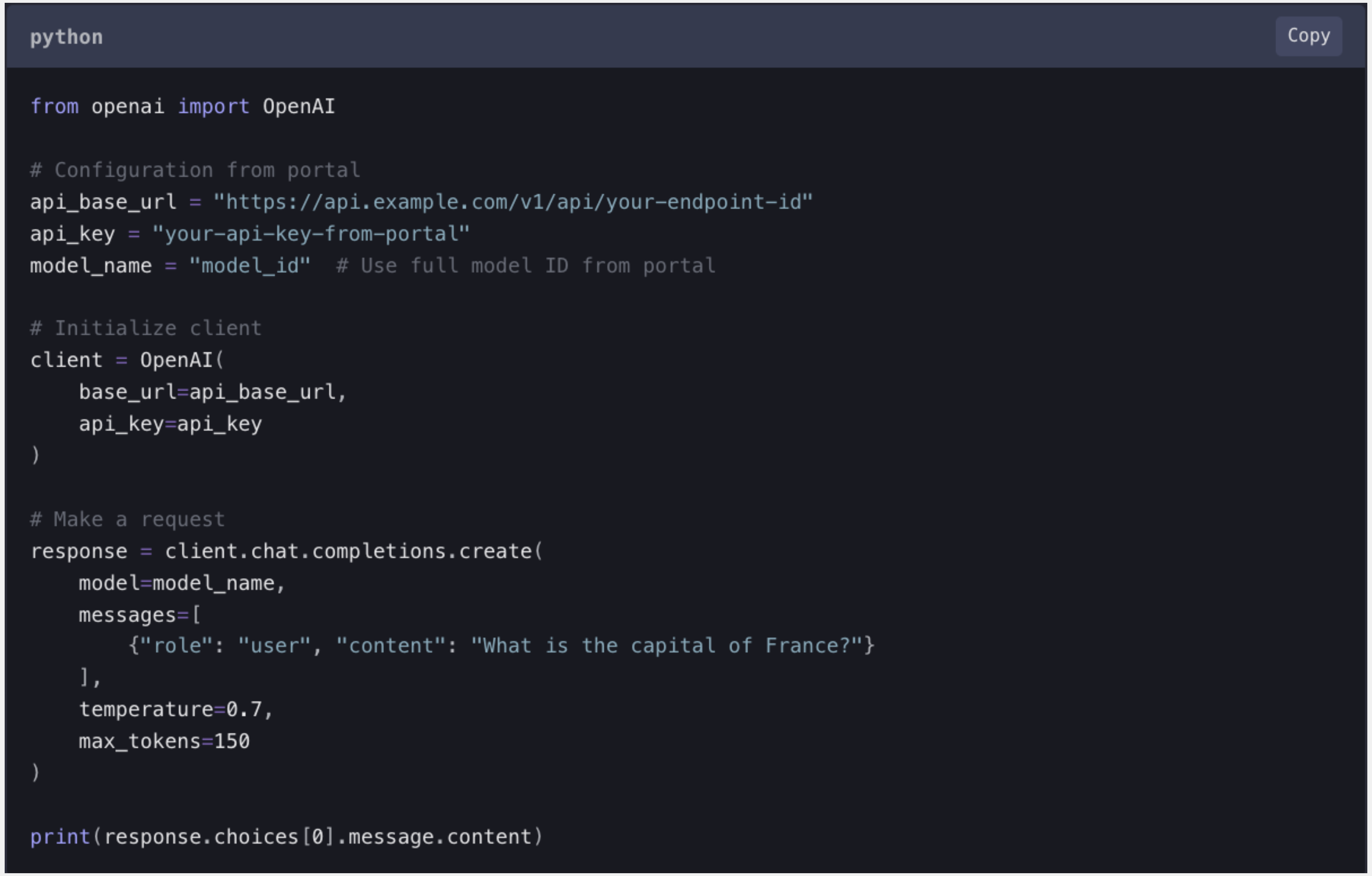
- Otel, Prometheus & ClickHouse based stack
- Token usage, latency, and GPU utilization metrics
- Per-tenant and per-service usage metrics
- End-to-end traces for RAG & agents
- Built-in billing and usage exports



CONFIDENTIAL Mind

API Compatibility & Integration

- OpenAI-compatible /v1/chat/completions
- Streaming responses
- Structured JSON schema support
- Multimodal support
- Compatible with existing AI frameworks & SDKs



```python
from openai import OpenAI

# Configuration from portal
api_base_url = "https://api.example.com/v1/api/your-endpoint-id"
api_key = "your-api-key-from-portal"
model_name = "model_id"  # Use full model ID from portal

# Initialize client
client = OpenAI(
    base_url=api_base_url,
    api_key=api_key
)

# Make a request
response = client.chat.completions.create(
    model=model_name,
    messages=[
        {"role": "user", "content": "What is the capital of France?"}
    ],
    temperature=0.7,
    max_tokens=150
)

print(response.choices[0].message.content)
```

Summary & Roadmap

Deployment Models

On-premises

Runs fully inside the customer’s own data center, giving complete control over data, security, and GPU infrastructure.

Private clouds & hyperscalers

Deploys into managed cloud environments (e.g., ElastX, Magic Cloud) while maintaining tenant isolation and enterprise-grade governance.

Public cloud

Allows organizations to leverage public cloud compute for rapid scaling today, while retaining the option to migrate and run fully on their own hardware in the future.

Edge & air-gapped

Operates in highly restricted, offline, or latency-sensitive environments with no external connectivity, suitable for defense, industrial, and remote setups.

Key Technical Differentiators

Enterprise-Grade Multi-Tenant AI

Multiple organizations and teams can safely share the same GPU infrastructure with strict tenant isolation and access control.

End-to-end governed AI platform

All AI usage is centrally managed with built-in security, auditing, access policies, and compliance controls from deployment to inference.

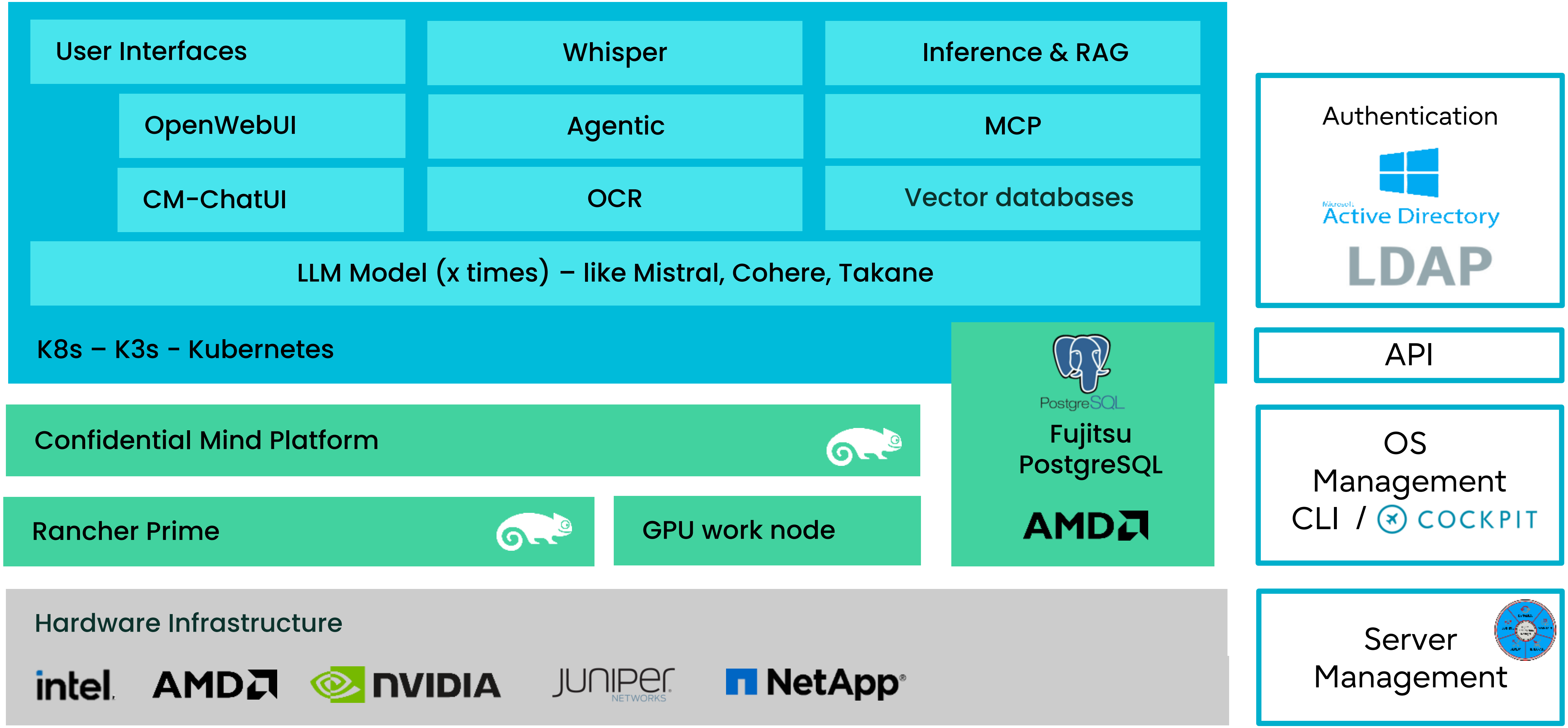
OpenAI-compatible AIaaS APIs

Standard OpenAI-style APIs enable drop-in compatibility with existing tools, SDKs, and applications without vendor lock-in.

All-in-One Platform for Models, RAG & Agents

LLMs, retrieval-augmented generation, and autonomous agents are all operated through a single integrated control plane.

Enterprise AI solution



User Interfaces

OpenWebUI

CM-ChatUI

Whisper

Agentic

OCR

Inference & RAG

MCP

Vector databases

LLM Model (x times) – like Mistral, Cohere, Takane

K8s – K3s - Kubernetes

PostgreSQL

Fujitsu PostgreSQL

AMD

Confidential Mind Platform

Rancher Prime

GPU work node

Hardware Infrastructure

intel AMD NVIDIA JUNIPER NETWORKS NetApp

Authentication

Microsoft Active Directory

LDAP

API

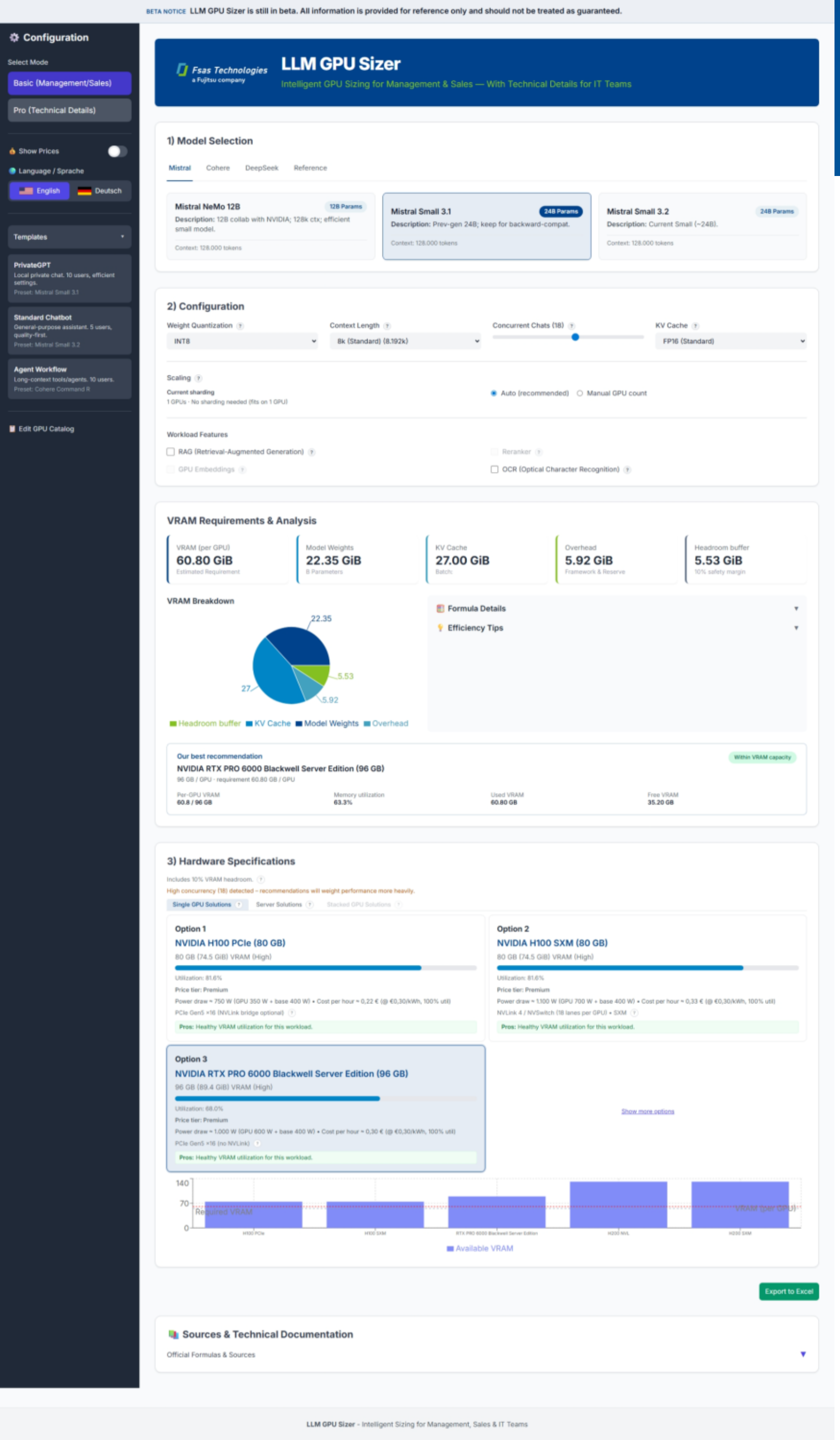
OS Management

CLI / COCKPIT

Server Management

LLM GPU sizer

Fsas Technologies
a Fujitsu company



BETA NOTICE LLM GPU Sizer is still in beta. All information is provided for reference only and should not be treated as guaranteed.

Configuration

Select Mode

Basic (Management/Sales)

Pro (Technical Details)

Show Prices

Language / Sprache

English | Deutsch

Templates ▾

PrivateGPT
Local private chat. 10 users, efficient settings.
Preset: Mistral Small 3.1

Standard Chatbot
General-purpose assistant. 5 users, quality-first.
Preset: Mistral Small 3.2

Agent Workflow
Long-context tools/agents. 10 users.
Preset: Cohere Command R

Edit GPU Catalog

LLM GPU Sizer
Intelligent GPU Sizing for Management & Sales — With Technical Details for IT Teams

1) Model Selection

Mistral | Cohere | DeepSeek | Reference

Mistral NeMo 12B — 12B Params
Description: 12B collab with NVIDIA; 128k ctx; efficient small model.
Context: 128.000 tokens

Mistral Small 3.1 — 24B Params
Description: Prev-gen 24B; keep for backward-compat.
Context: 128.000 tokens

Mistral Small 3.2 — 24B Params
Description: Current Small (~24B).
Context: 128.000 tokens

2) Configuration

Weight Quantization: INT8
Context Length: 8k (Standard) (8.192k)
Concurrent Chats (18)
KV Cache: FP16 (Standard)

Scaling
Current sharding
1 GPUs · No sharding needed (fits on 1 GPU)
● Auto (recommended) ○ Manual GPU count

Workload Features
☐ RAG (Retrieval-Augmented Generation)
☐ Reranker
☐ GPU Embeddings
☐ OCR (Optical Character Recognition)

VRAM Requirements & Analysis

VRAM (per GPU)	Model Weights	KV Cache	Overhead	Headroom buffer
60.80 GiB	22.35 GiB	27.00 GiB	5.92 GiB	5.53 GiB
Estimated Requirement	B Parameters	Batch:	Framework & Reserve	10% safety margin

VRAM Breakdown

Formula Details ▾
Efficiency Tips ▾

Our best recommendation — Within VRAM capacity
NVIDIA RTX PRO 6000 Blackwell Server Edition (96 GB)
96 GB / GPU · requirement 60.80 GB / GPU

Per-GPU VRAM	Memory utilization	Used VRAM	Free VRAM
60.8 / 96 GB	63.3%	60.80 GB	35.20 GB

3) Hardware Specifications

Includes 10% VRAM headroom.
High concurrency (18) detected – recommendations will weight performance more heavily.

Single GPU Solutions | Server Solutions | Stacked GPU Solutions

Option 1
NVIDIA H100 PCIe (80 GB)
80 GB (74.5 GiB) VRAM (High)
Utilization: 81.6%
Price tier: Premium
Power draw ≈ 750 W (GPU 350 W + base 400 W) • Cost per hour ≈ 0,22 € (@ €0,30/kWh, 100% util)
PCIe Gen5 ×16 (NVLink bridge optional)
Pros: Healthy VRAM utilization for this workload.

Option 2
NVIDIA H100 SXM (80 GB)
80 GB (74.5 GiB) VRAM (High)
Utilization: 81.6%
Price tier: Premium
Power draw ≈ 1.100 W (GPU 700 W + base 400 W) • Cost per hour ≈ 0,33 € (@ €0,30/kWh, 100% util)
NVLink 4 / NVSwitch (18 lanes per GPU) • SXM
Pros: Healthy VRAM utilization for this workload.

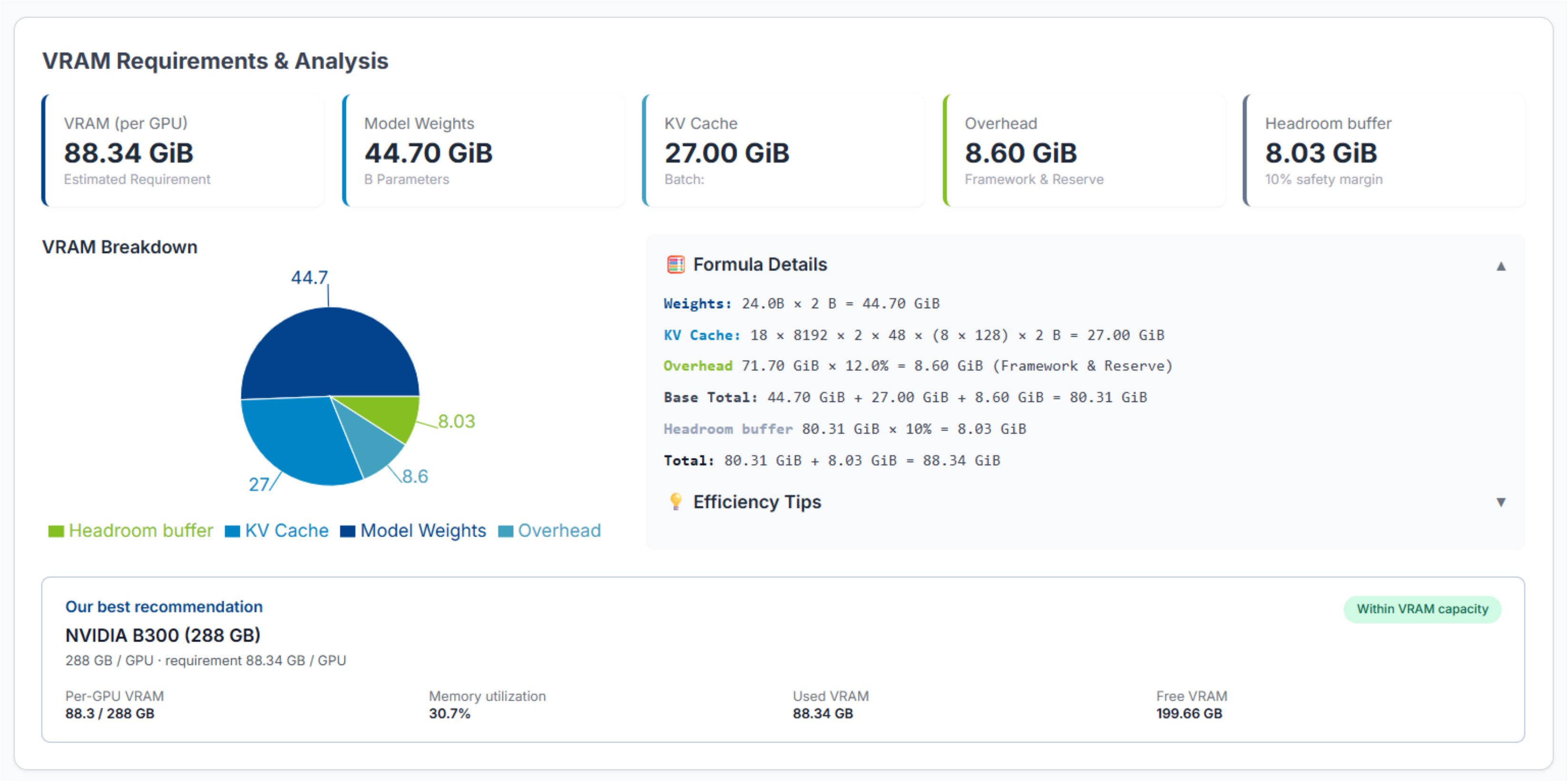
Option 3
NVIDIA RTX PRO 6000 Blackwell Server Edition (96 GB)
96 GB (89.4 GiB) VRAM (High)
Utilization: 68.0%
Price tier: Premium
Power draw ≈ 1.000 W (GPU 600 W + base 400 W) • Cost per hour ≈ 0,30 € (@ €0,30/kWh, 100% util)
PCIe Gen5 ×16 (no NVLink)
Pros: Healthy VRAM utilization for this workload.

Show more options

Export to Excel

Sources & Technical Documentation
Official Formulas & Sources ▾

LLM GPU Sizer - Intelligent Sizing for Management, Sales & IT Teams



VRAM Requirements & Analysis

VRAM (per GPU)	Model Weights	KV Cache	Overhead	Headroom buffer
88.34 GiB	44.70 GiB	27.00 GiB	8.60 GiB	8.03 GiB
Estimated Requirement	B Parameters	Batch:	Framework & Reserve	10% safety margin

VRAM Breakdown

Formula Details ▴

Weights: 24.0B × 2 B = 44.70 GiB
KV Cache: 18 × 8192 × 2 × 48 × (8 × 128) × 2 B = 27.00 GiB
Overhead 71.70 GiB × 12.0% = 8.60 GiB (Framework & Reserve)
Base Total: 44.70 GiB + 27.00 GiB + 8.60 GiB = 80.31 GiB
Headroom buffer 80.31 GiB × 10% = 8.03 GiB
Total: 80.31 GiB + 8.03 GiB = 88.34 GiB

Efficiency Tips ▾

Our best recommendation — Within VRAM capacity
NVIDIA B300 (288 GB)
288 GB / GPU · requirement 88.34 GB / GPU

Per-GPU VRAM	Memory utilization	Used VRAM	Free VRAM
88.3 / 288 GB	30.7%	88.34 GB	199.66 GB

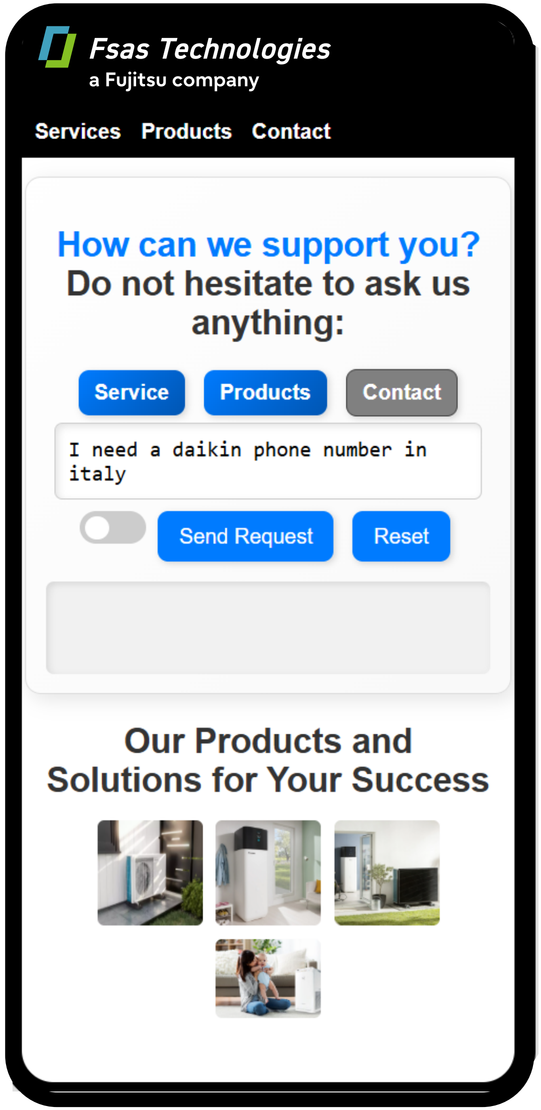
http://r01azws01twahe.r01.fujitsu.local/gpusizer/index.html

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

UseCases
Anwendungsfälle

Chatbot Endpoint: Additonal Usage

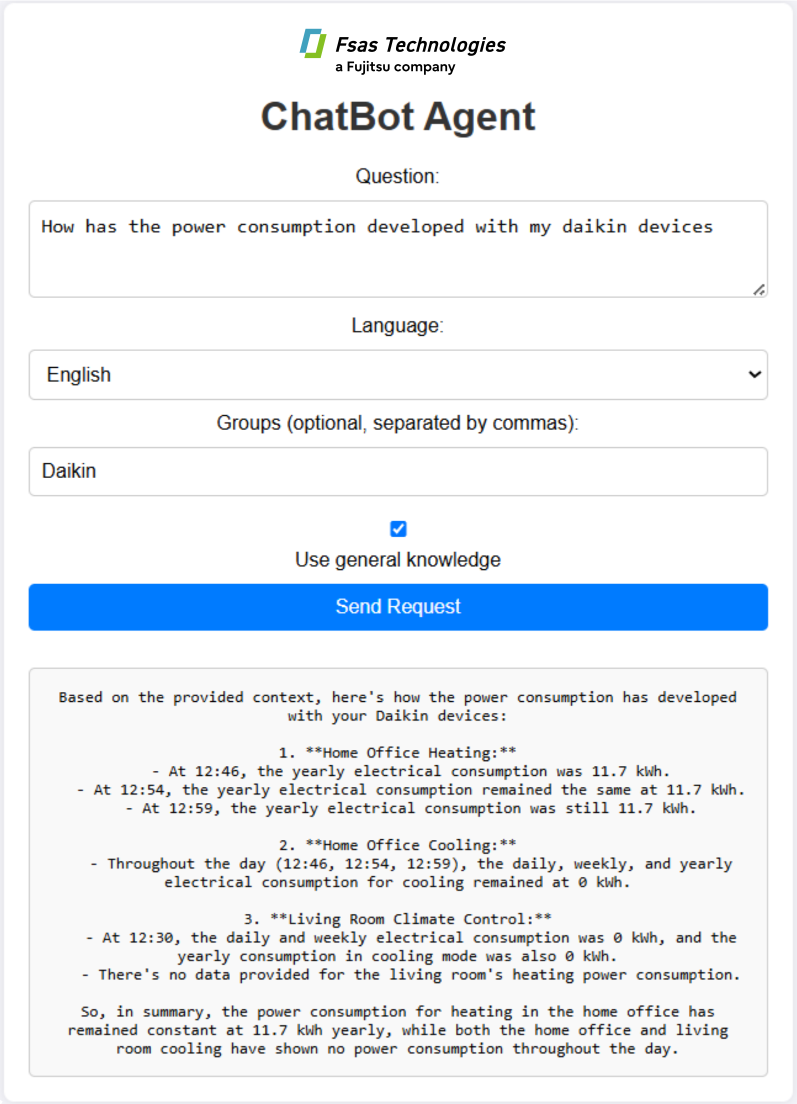
Fsas Technologies
a Fujitsu company



How can we support you? Do not hesitate to ask us anything:

I need a daikin phone number in italy

Our Products and Solutions for Your Success



ChatBot Agent

Question: How has the power consumption developed with my daikin devices

Language: English

Groups (optional, separated by commas): Daikin

☑ Use general knowledge

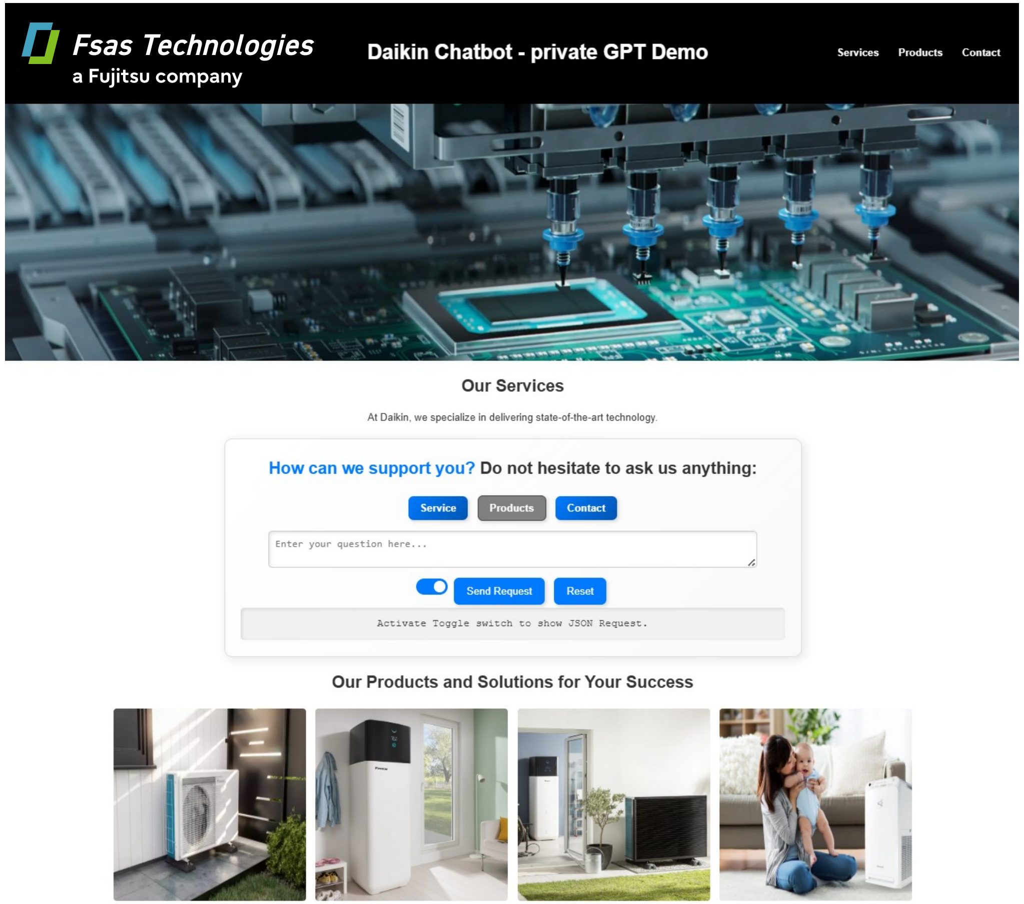
```
Based on the provided context, here's how the power consumption has developed
with your Daikin devices:

1. **Home Office Heating:**
- At 12:46, the yearly electrical consumption was 11.7 kWh.
- At 12:54, the yearly electrical consumption remained the same at 11.7 kWh.
- At 12:59, the yearly electrical consumption was still 11.7 kWh.

2. **Home Office Cooling:**
- Throughout the day (12:46, 12:54, 12:59), the daily, weekly, and yearly
electrical consumption for cooling remained at 0 kWh.

3. **Living Room Climate Control:**
- At 12:30, the daily and weekly electrical consumption was 0 kWh, and the
yearly consumption in cooling mode was also 0 kWh.
- There's no data provided for the living room's heating power consumption.

So, in summary, the power consumption for heating in the home office has
remained constant at 11.7 kWh yearly, while both the home office and living
room cooling have shown no power consumption throughout the day.
```



Our Services

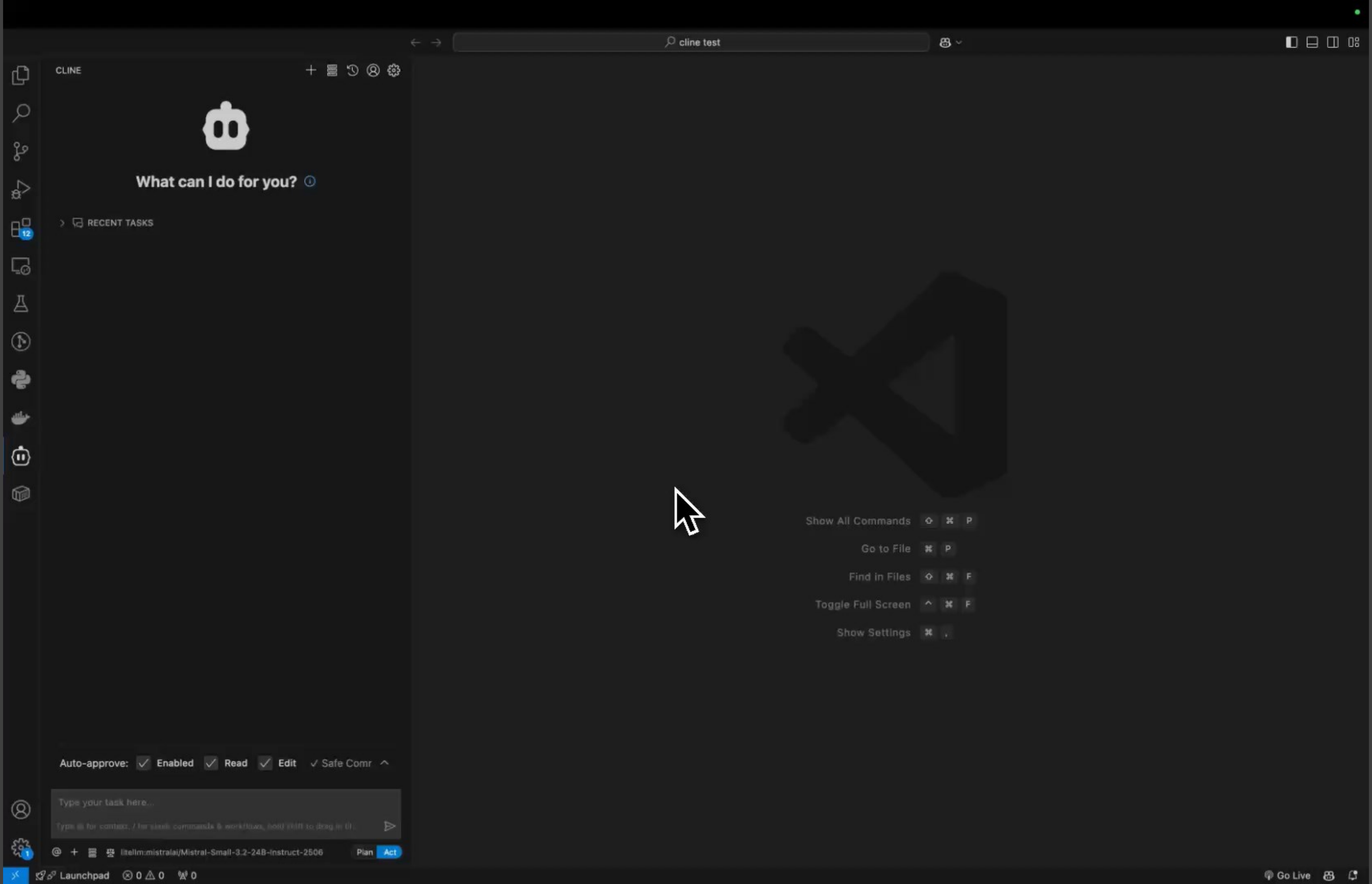
At Daikin, we specialize in delivering state-of-the-art technology.

How can we support you? Do not hesitate to ask us anything:

Activate Toggle switch to show JSON Request.

Our Products and Solutions for Your Success

© 2025 Fsas Technologies



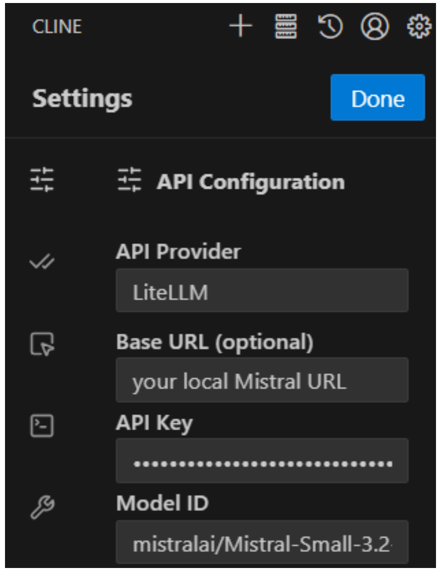
Auto-approve: ☑ Enabled ☑ Read ☑ Edit ✓ Safe Comr

litellm:mistralai/Mistral-Small-3.2-24B-Instruct-2506

Agentic Coding

Mistral Small + Cline in VSCode

Description: We are using the on premise LLM Mistral Small 3.2 together with the Agentic Coding capabilities of Cline to build a Hardware Compatibility Configurator in one shot.

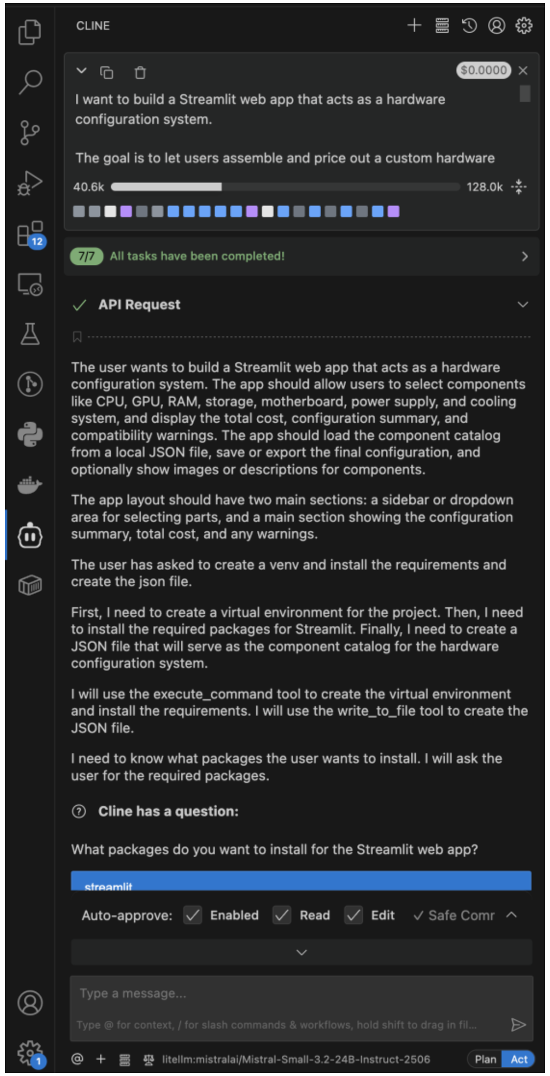


API Provider: LiteLLM

Base URL (optional): your local Mistral URL

API Key: ••••••••••••••••••••••••••••

Model ID: mistralai/Mistral-Small-3.2



I want to build a Streamlit web app that acts as a hardware configuration system.

The goal is to let users assemble and price out a custom hardware

The user wants to build a Streamlit web app that acts as a hardware configuration system. The app should allow users to select components like CPU, GPU, RAM, storage, motherboard, power supply, and cooling system, and display the total cost, configuration summary, and compatibility warnings. The app should load the component catalog from a local JSON file, save or export the final configuration, and optionally show images or descriptions for components.

The app layout should have two main sections: a sidebar or dropdown area for selecting parts, and a main section showing the configuration summary, total cost, and any warnings.

The user has asked to create a venv and install the requirements and create the json file.

First, I need to create a virtual environment for the project. Then, I need to install the required packages for Streamlit. Finally, I need to create a JSON file that will serve as the component catalog for the hardware configuration system.

I will use the execute_command tool to create the virtual environment and install the requirements. I will use the write_to_file tool to create the JSON file.

I need to know what packages the user wants to install. I will ask the user for the required packages.

Cline has a question:

What packages do you want to install for the Streamlit web app?

streamlit

Auto-approve: ☑ Enabled ☑ Read ☑ Edit ✓ Safe Comr

Type a message...

Type @ for context, / for slash commands & workflows, hold shift to drag in fil...

litellm:mistralai/Mistral-Small-3.2-24B-Instruct-2506

Core Capabilties:

- Create, edit, and modify files across your entire project
- Execute terminal commands (with your approval)
- Analyze codebases and understand project structure
- Debug issues and fix errors automatically
- Run in Plan & Act modes for strategic thinking vs implementation

Privacy & Control :

- Complete privacy - your code never leaves your machine
- Zero API costs after initial setup
- Work completely offline - no internet dependency
- Full control over model versions and configurations
- No vendor lock-in or subscription fees

Key Use Cases:

- Sensitive/proprietary code development
- Air-gapped environments
- Learning and experimentation without costs
- High-volume development work
- Remote locations with poor connectivity
- Compliance-required projects

Setup

1. Install Cline VS Code Extension
2. Insert Local Mistral Url + Key

Start Coding!

n8n and PrivateGPT a successful duo

A decision tree plus LLM workflow that keeps tickets complete, assigned, and up-to-date.

Before

- Incomplete fields
- Unassigned tasks → no accountability
- Comment histories over weeks → slow context build-up

Runs fully on-prem for data sovereignty (n8n + On-Prem-LLM)

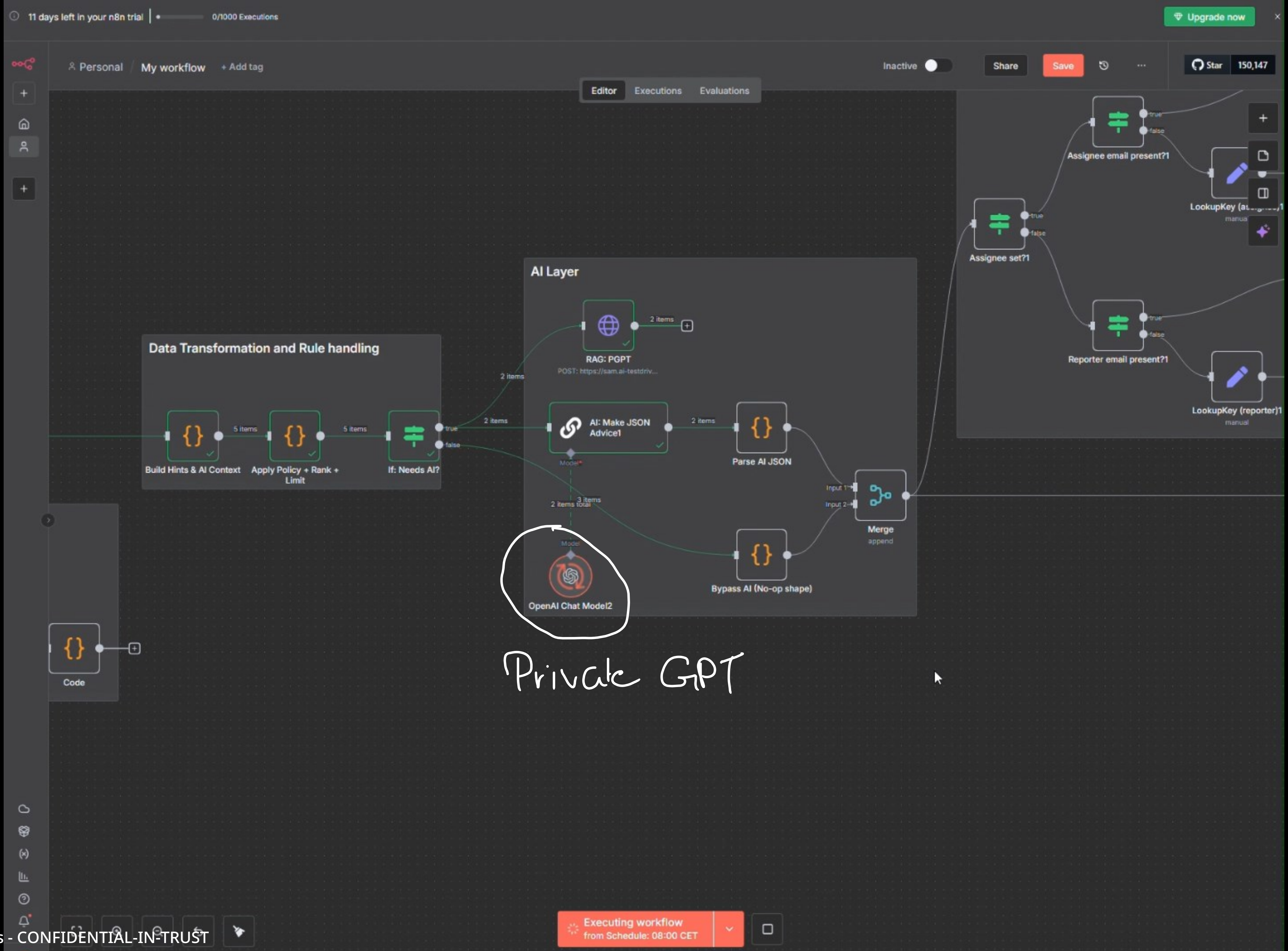
After

- Auto-validation on creation
- Weekly run checks rules and flags stale tickets
- AI briefing: current status + next best steps, delivered via email.

n8n

PGPT

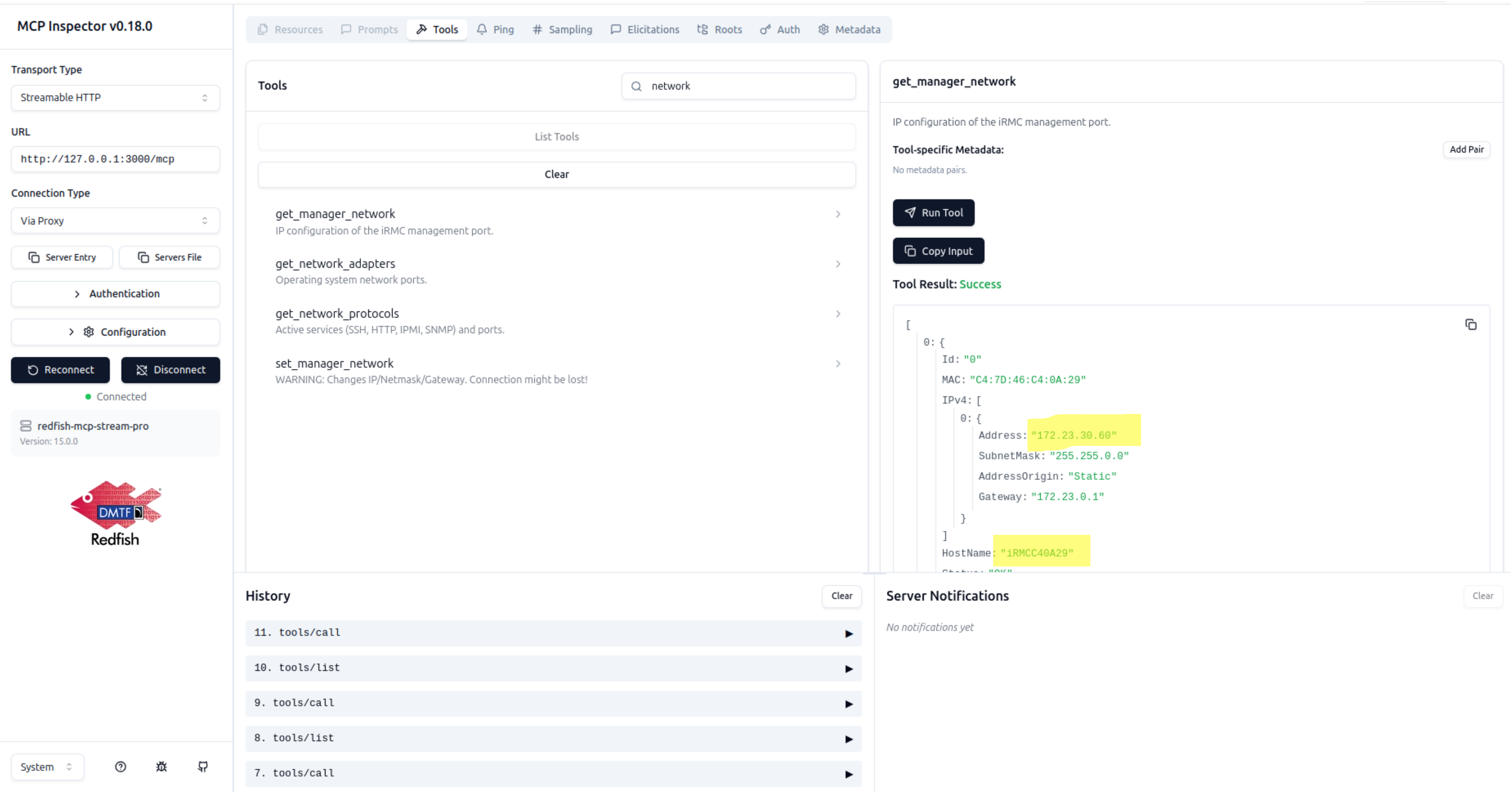
Orchestrierung | *Infra* | *LLM*



Private GPT

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

PrivateGPT in the DC with REDfish via MCP (III)



Transport Type: Streamable HTTP

URL: `http://127.0.0.1:3000/mcp`

Connection Type: Via Proxy

Connected

redfish-mcp-stream-pro
Version: 15.0.0

Tools (search: network)

- get_manager_network — IP configuration of the iRMC management port.
- get_network_adapters — Operating system network ports.
- get_network_protocols — Active services (SSH, HTTP, IPMI, SNMP) and ports.
- set_manager_network — WARNING: Changes IP/Netmask/Gateway. Connection might be lost!

get_manager_network

IP configuration of the iRMC management port.

Tool-specific Metadata:

No metadata pairs.

Tool Result: Success

```
[
  0: {
    Id: "0"
    MAC: "C4:7D:46:C4:0A:29"
    IPv4: [
      0: {
        Address: "172.23.30.60"
        SubnetMask: "255.255.0.0"
        AddressOrigin: "Static"
        Gateway: "172.23.0.1"
      }
    ]
    HostName: "iRMCC40A29"
```

History

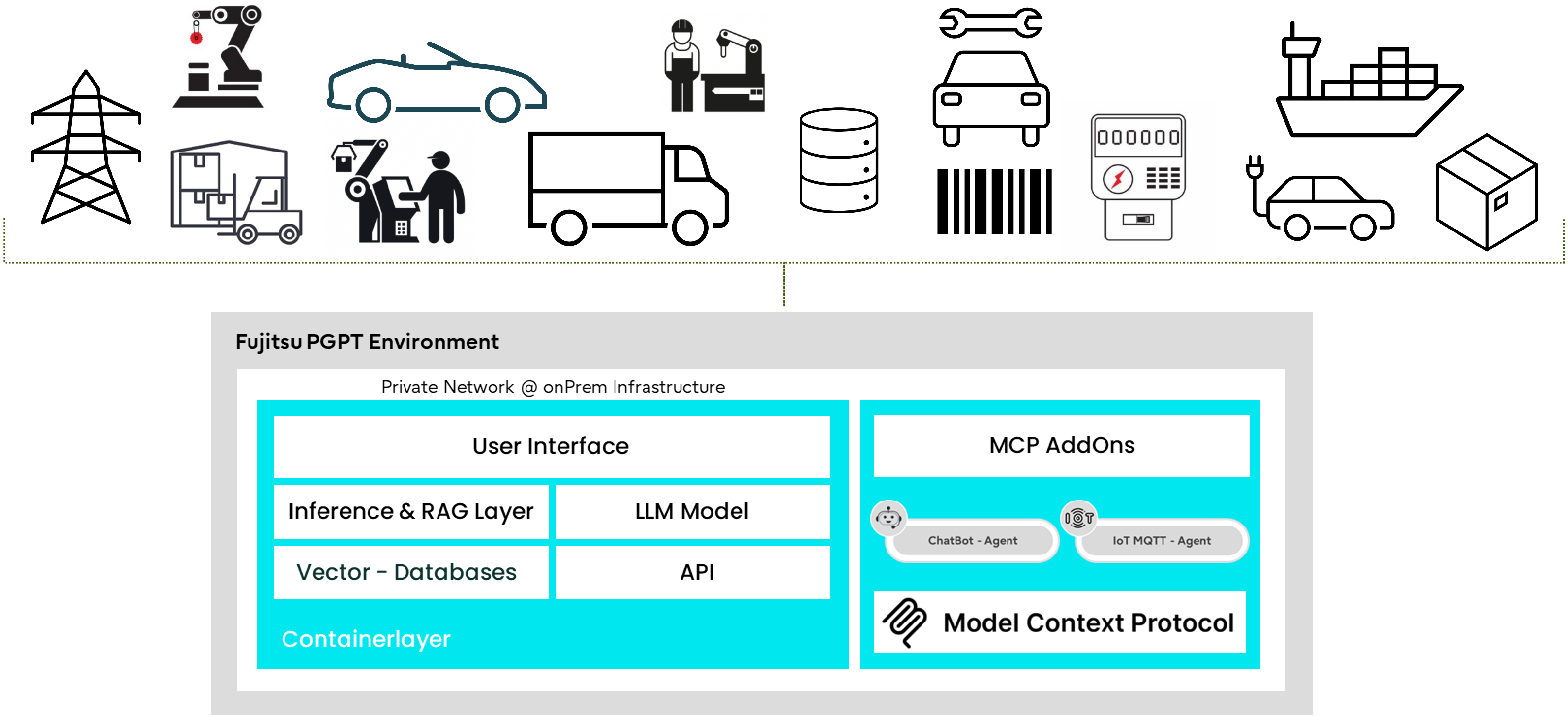
```
11. tools/call
10. tools/list
9. tools/call
8. tools/list
7. tools/call
```

Server Notifications

No notifications yet

© 2026 Fsas Technologies - CONFIDENTIAL-IN-TRUST

IoT meets LLM



Fujitsu PGPT Environment

Private Network @ onPrem Infrastructure

User Interface

Inference & RAG Layer

LLM Model

Vector - Databases

API

Containerlayer

MCP AddOns

ChatBot - Agent

IoT MQTT - Agent

Model Context Protocol

Office of the CTO

MAMORU

AI based Server Management

Jochen Riedisser
CDO, Fsas EPB
Fujitsu Fellow

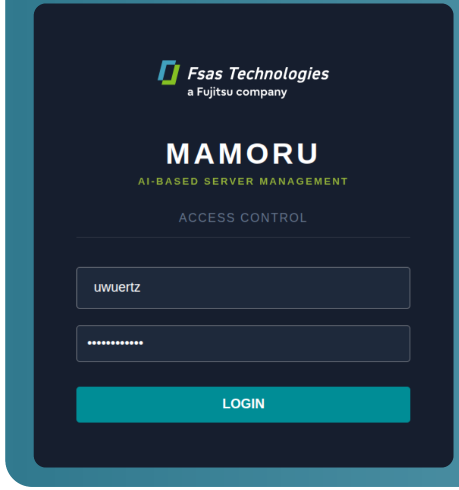
FUJITSU-CONFIDENTIAL-IN-TRUST

MAMORU Technology - Stack

MAMORU CLI

CLI

MAMORU WEB UI



Fsas Technologies
a Fujitsu company

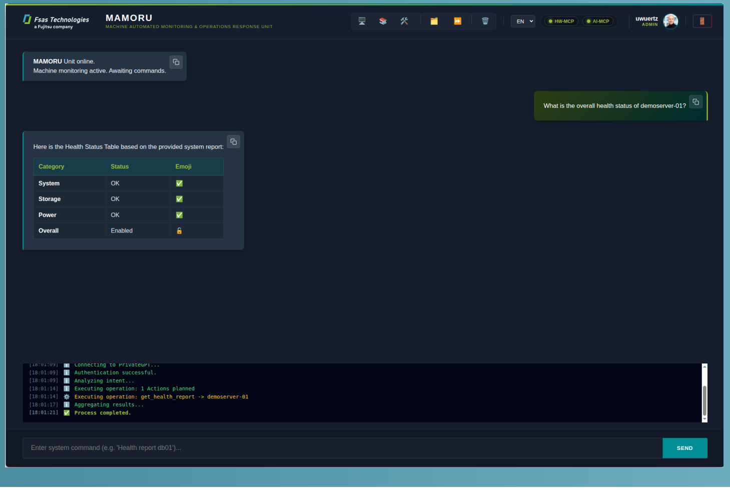
MAMORU

AI-BASED SERVER MANAGEMENT

ACCESS CONTROL

uwuertz

••••••••••••



MAMORU Unit online.
Machine monitoring active. Awaiting commands.

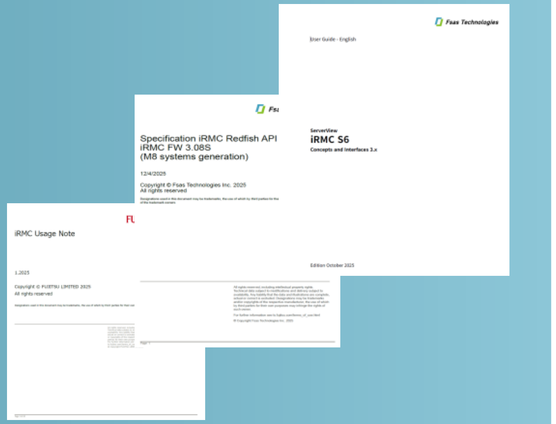
What is the overall health status of demoserver-01?

Here is the Health Status Table based on the provided system report:

Category	Status	Emoji
System	OK	✅
Storage	OK	✅
Power	OK	✅
Overall	Enabled	🔓

```
[18:01:09] ℹ️ Connecting to PrivateGPT...
[18:01:09] ℹ️ Authentication successful.
[18:01:09] ℹ️ Analyzing intent...
[18:01:14] ℹ️ Executing operation: 1 Actions planned
[18:01:14] ⚙️ Executing operation: get_health_report -> demoserver-01
[18:01:17] ℹ️ Aggregating results...
[18:01:21] ✅ Process completed.
```

Enter system command (e.g. 'Health report db01')...



Fsas Technologies
User Guide - English
ServerView
iRMC S6
Concepts and Interfaces 3.x
Edition October 2025

Specification iRMC Redfish API
iRMC FW 3.08S
(M8 systems generation)
12/4/2025
Copyright © Fsas Technologies Inc. 2025
All rights reserved

iRMC Usage Note
1.2025
Copyright © FUJITSU LIMITED 2025
All rights reserved

MAMORU AI Agent – Logical Layer

MCP Server Redfish
Physical Connector

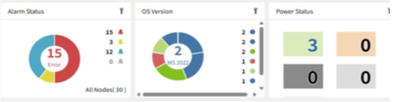
MCP Server ISM
Management Connector

MCP Server PGPT
Intelligence Connector

Manuals & Operation Guides via RAG

IT-Infrastructure

ISM (Optional)



Private GPT

RAG

Takane

MISTRAL AI_

© 2025 Fsas Technologies

Redfish 3.08S Implementation, more than 60 Functions via MCP

A Global Standard for Server Management

Office of the CTO

Function	Description
clear_system_logs	Clears all entries in the System Event Log (SEL).
configure_cas	Configures Central Authentication Service (CAS).
configure_ldap_full	Comprehensive LDAP configuration (Primary, Backup, Directory Settings).
configure_smtp	Configures the SMTP mail server for alerts.
create_event_subscription	Subscribes to server events (Webhook/SNMP).
toggle_uid_led	ID-LED control (Chassis).
trigger_firmware_update	Starts a firmware update via URI (SimpleUpdate).
trigger_test_alert	Sends a test alert (Check SNMP/Mail).
get_storage_info	RAID controllers and drives.
get_system_logs	Smart Search Logs (SEL/IEL).
get_system_summary	Basic info: Serial, Model, AssetTag, Power, RAM, CPU.
get_tasks	Running background tasks.
	•••

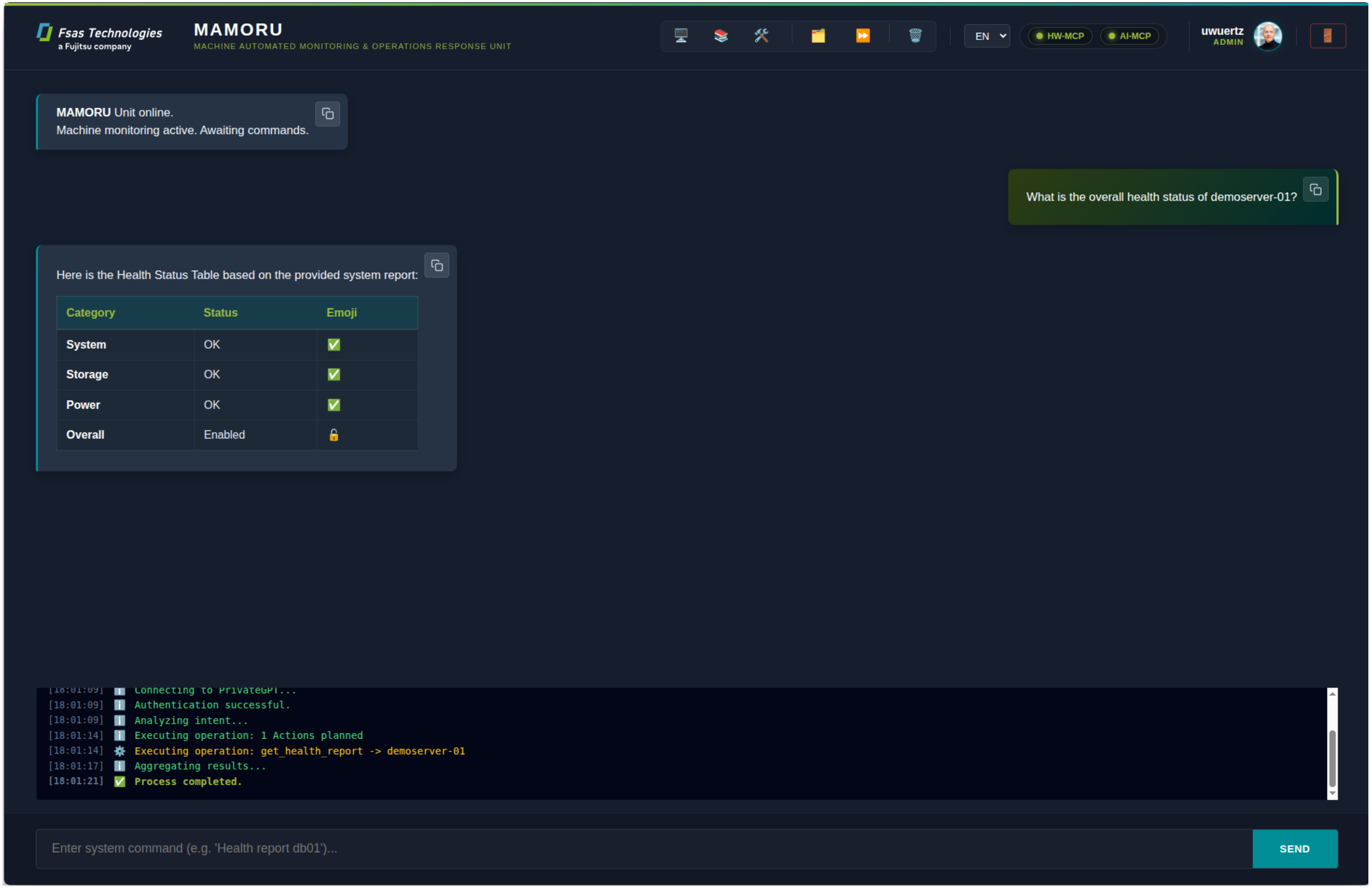
Function	Description
get_health_report	Overall status (Health Rollup) for System, Storage, Power.
get_license_info	Licenses (incl. Fujitsu OEM).
get_manager_network	IP configuration of the iRMC management port.
get_network_adapters	Operating system network ports.
get_network_protocols	Active services (SSH, HTTP, IPMI, SNMP) and ports.
create_raid_volume	WARNING: Creates a RAID volume. Data loss possible.
create_user_account	Creates a new user account.
delete_raid_volume	DANGER: Deletes a RAID volume and ALL DATA.
delete_user_account	Deletes a user via ID.
get_thermal_data	Temperatures and fans.
get_user_accounts	Lists iRMC users.
identify_drive	Blinks the LED of a hard drive (Locate).
	•••

© 2025 Fsas Technologies

MAMORU Screenshots (Examples)

Reliable Information at your Fingertips, Fast & Easy

Office of the CTO



MAMORU Unit online.
Machine monitoring active. Awaiting commands.

What is the overall health status of demoserver-01?

Here is the Health Status Table based on the provided system report:

Category	Status	Emoji
System	OK	✅
Storage	OK	✅
Power	OK	✅
Overall	Enabled	🔓

```
[18:01:09] ℹ️ Connecting to PrivateGPT...
[18:01:09] ℹ️ Authentication successful.
[18:01:09] ℹ️ Analyzing intent...
[18:01:14] ℹ️ Executing operation: 1 Actions planned
[18:01:14] ⚙️ Executing operation: get_health_report -> demoserver-01
[18:01:17] ℹ️ Aggregating results...
[18:01:21] ✅ Process completed.
```

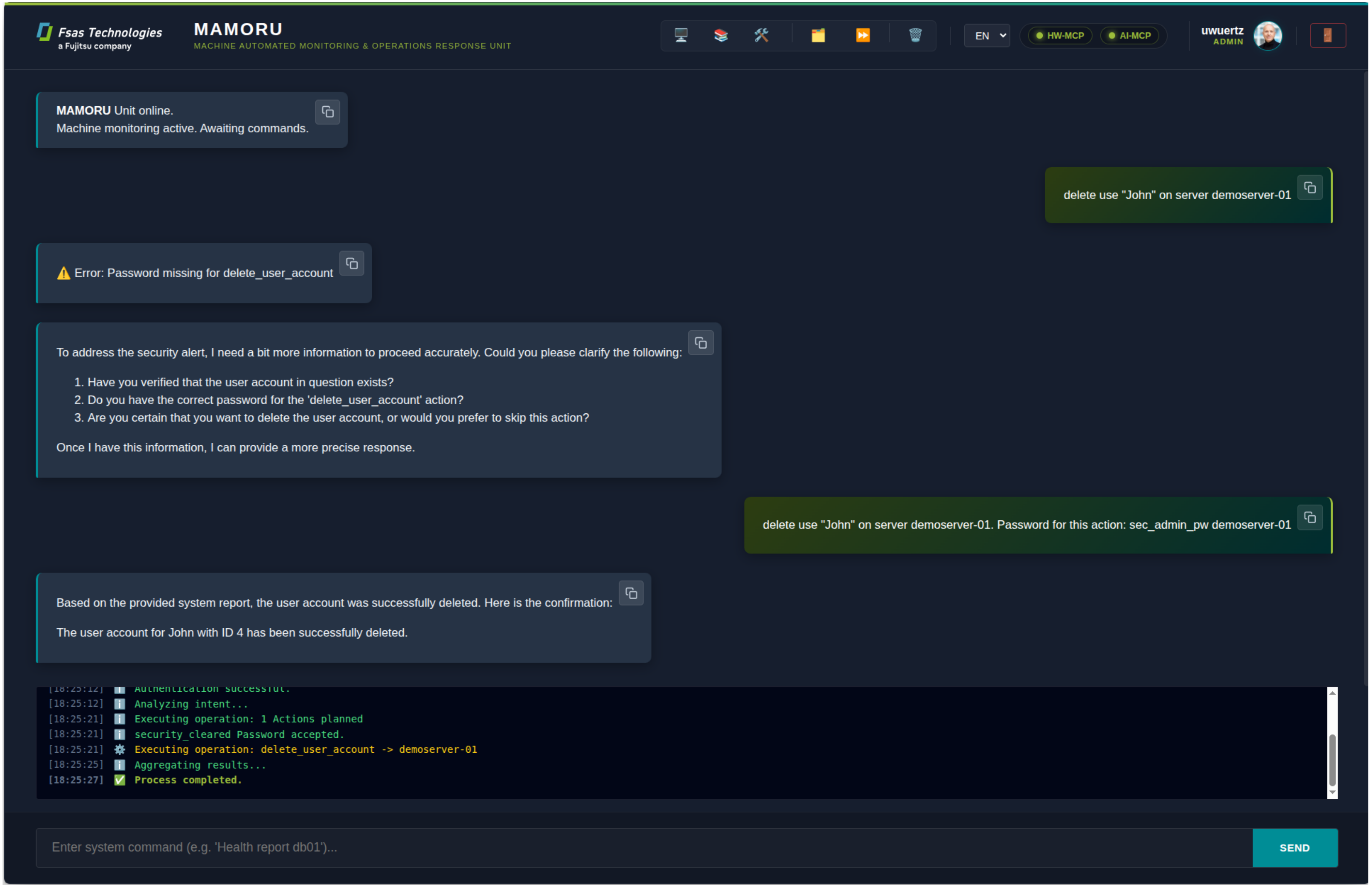
Enter system command (e.g. 'Health report db01')...

© 2025 Fsas Technologies

MAMORU Screenshots (Examples)

Additional approval password required for critical LLM operations (4-Eyes Principle)

Office of the CTO



MAMORU Unit online.
Machine monitoring active. Awaiting commands.

delete use "John" on server demoserver-01

⚠️ Error: Password missing for delete_user_account

To address the security alert, I need a bit more information to proceed accurately. Could you please clarify the following:

1. Have you verified that the user account in question exists?
2. Do you have the correct password for the 'delete_user_account' action?
3. Are you certain that you want to delete the user account, or would you prefer to skip this action?

Once I have this information, I can provide a more precise response.

delete use "John" on server demoserver-01. Password for this action: sec_admin_pw demoserver-01

Based on the provided system report, the user account was successfully deleted. Here is the confirmation:

The user account for John with ID 4 has been successfully deleted.

```
[18:25:12] ℹ️ Authentication successful.
[18:25:12] ℹ️ Analyzing intent...
[18:25:21] ℹ️ Executing operation: 1 Actions planned
[18:25:21] ℹ️ security_cleared Password accepted.
[18:25:21] ⚙️ Executing operation: delete_user_account -> demoserver-01
[18:25:25] ℹ️ Aggregating results...
[18:25:27] ✅ Process completed.
```

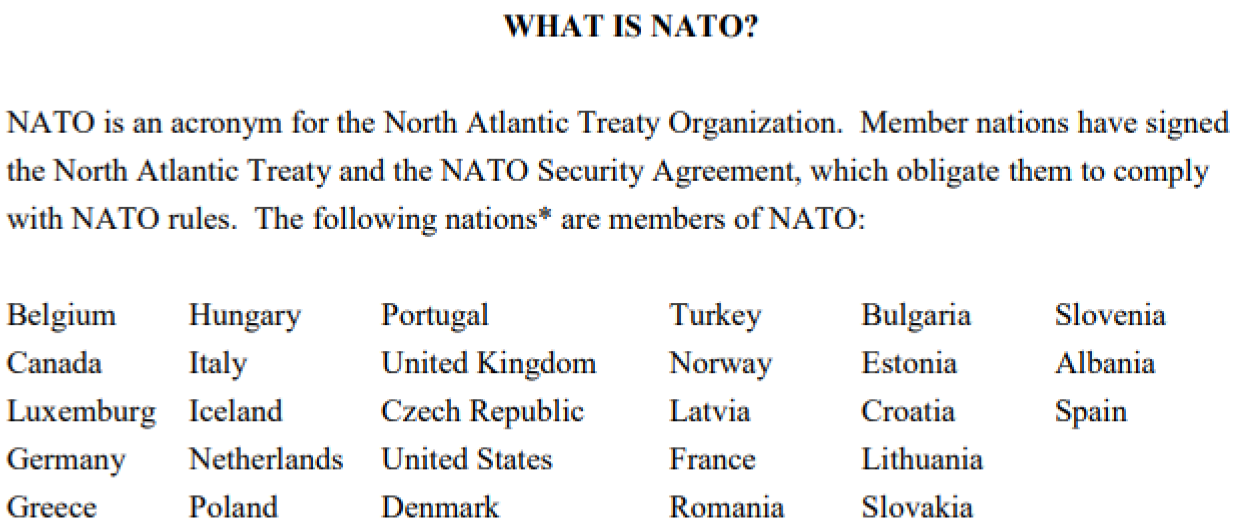
© 2025 Fsas Technologies

Defense

Fsas PrivateGPT and MCP Use Cases

Sovereign AI for Defense Operations

NATO members / classification (internal / remove slide)



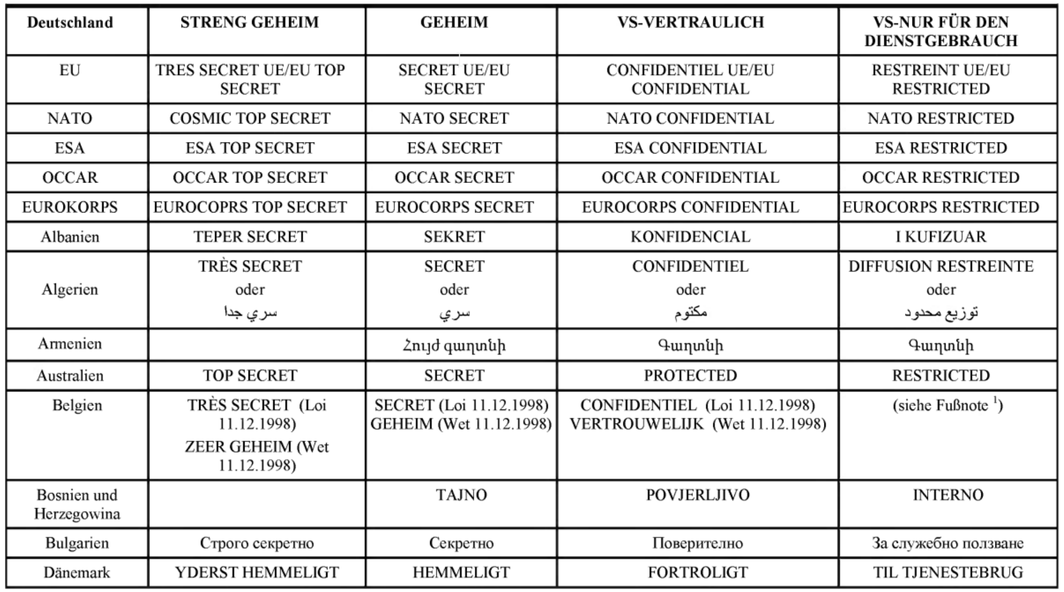
WHAT IS NATO?

NATO is an acronym for the North Atlantic Treaty Organization. Member nations have signed the North Atlantic Treaty and the NATO Security Agreement, which obligate them to comply with NATO rules. The following nations* are members of NATO:

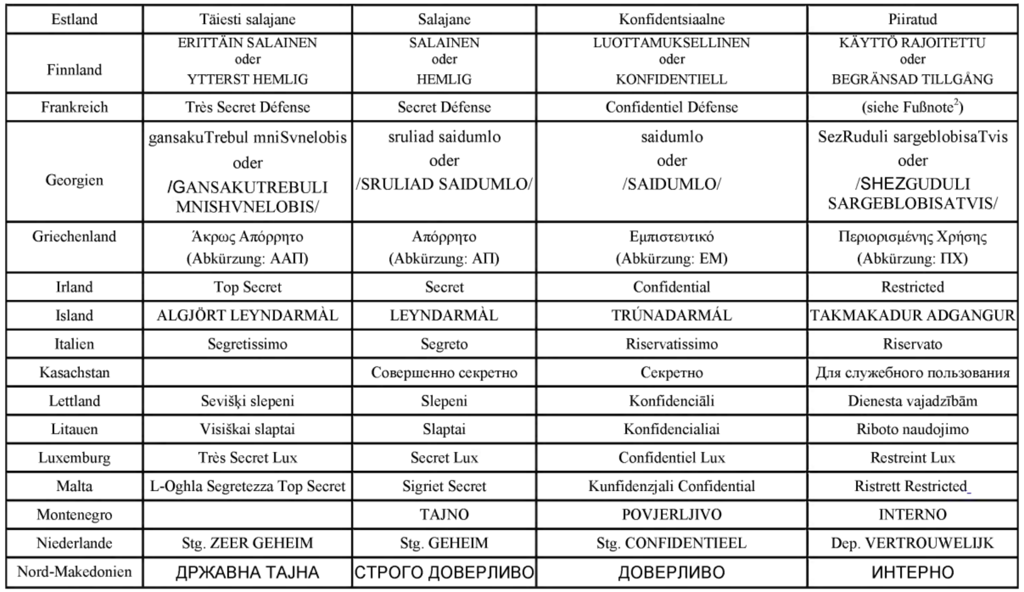
Belgium
Canada
Luxemburg
Germany
Greece
Hungary
Italy
Iceland
Netherlands
Poland
Portugal
United Kingdom
Czech Republic
United States
Denmark
Turkey
Norway
Latvia
France
Romania
Bulgaria
Estonia
Croatia
Lithuania
Slovakia
Slovenia
Albania
Spain

NATO classification also for use cases (internal / remove slide)

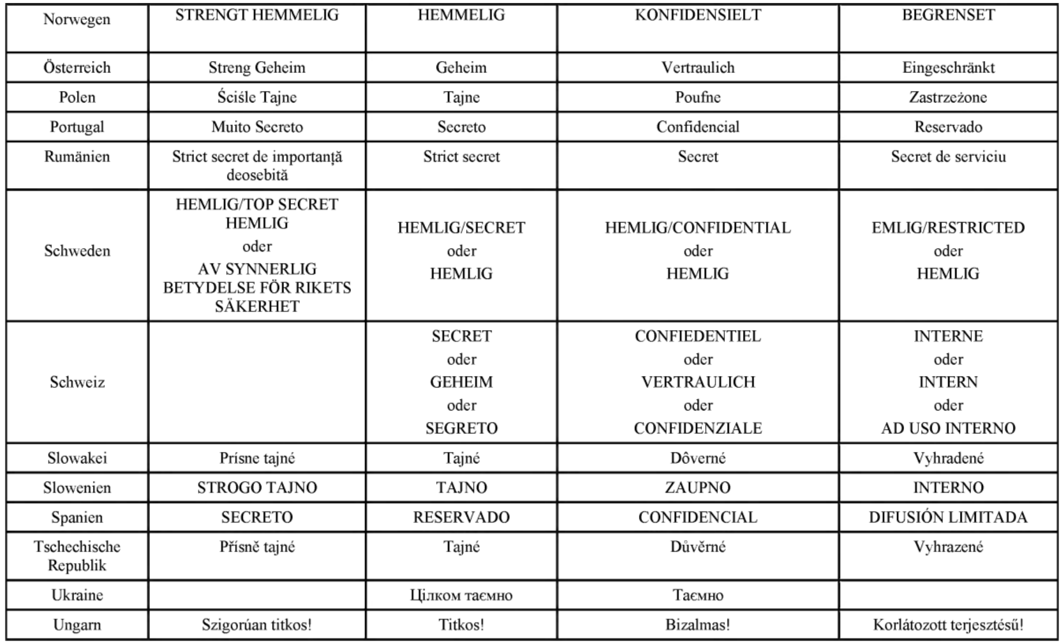
Fsas Technologies
a Fujitsu company



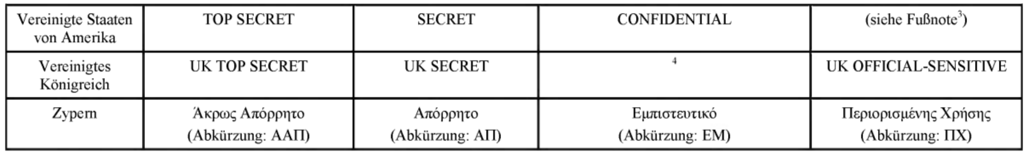
Deutschland	STRENG GEHEIM	GEHEIM	VS-VERTRAULICH	VS-NUR FÜR DEN DIENSTGEBRAUCH
EU	TRES SECRET UE/EU TOP SECRET	SECRET UE/EU SECRET	CONFIDENTIEL UE/EU CONFIDENTIAL	RESTREINT UE/EU RESTRICTED
NATO	COSMIC TOP SECRET	NATO SECRET	NATO CONFIDENTIAL	NATO RESTRICTED
ESA	ESA TOP SECRET	ESA SECRET	ESA CONFIDENTIAL	ESA RESTRICTED
OCCAR	OCCAR TOP SECRET	OCCAR SECRET	OCCAR CONFIDENTIAL	OCCAR RESTRICTED
EUROKORPS	EUROCOPRS TOP SECRET	EUROCORPS SECRET	EUROCORPS CONFIDENTIAL	EUROCORPS RESTRICTED
Albanien	TEPER SECRET	SEKRET	KONFIDENCIAL	I KUFIZUAR
Algerien	TRÈS SECRET oder سري جدا	SECRET oder سري	CONFIDENTIEL oder مكتوم	DIFFUSION RESTREINTE oder توزيع محدود
Armenien		Հույժ գաղտնի	Գաղտնի	Գաղտնի
Australien	TOP SECRET	SECRET	PROTECTED	RESTRICTED
Belgien	TRÈS SECRET (Loi 11.12.1998) ZEER GEHEIM (Wet 11.12.1998)	SECRET (Loi 11.12.1998) GEHEIM (Wet 11.12.1998)	CONFIDENTIEL (Loi 11.12.1998) VERTROUWELIJK (Wet 11.12.1998)	(siehe Fußnote [1])
Bosnien und Herzegowina		TAJNO	POVJERLJIVO	INTERNO
Bulgarien	Строго секретно	Секретно	Поверително	За служебно ползване
Dänemark	YDERST HEMMELIGT	HEMMELIGT	FORTROLIGT	TIL TJENESTEBRUG



Estland	Täiesti salajane	Salajane	Konfidentsiaalne	Piiratud
Finnland	ERITTÄIN SALAINEN oder YTTERST HEMLIG	SALAINEN oder HEMLIG	LUOTTAMUKSELLINEN oder KONFIDENTIELL	KÄYTTÖ RAJOITETTU oder BEGRÄNSAD TILLGÅNG
Frankreich	Très Secret Défense	Secret Défense	Confidentiel Défense	(siehe Fußnote[2])
Georgien	gansakuTrebul mniSvnelobis oder /GANSAKUTREBULI MNISHVNELOBIS/	sruliad saidumlo oder /SRULIAD SAIDUMLO/	saidumlo oder /SAIDUMLO/	SezRuduli sargeblobisaTvis oder /SHEZGUDULI SARGEBLOBISATVIS/
Griechenland	Άκρως Απόρρητο (Abkürzung: ΑΑΠ)	Απόρρητο (Abkürzung: ΑΠ)	Εμπιστευτικό (Abkürzung: ΕΜ)	Περιορισμένης Χρήσης (Abkürzung: ΠΧ)
Irland	Top Secret	Secret	Confidential	Restricted
Island	ALGJÖRT LEYNDARMÀL	LEYNDARMÀL	TRÚNADARMÁL	TAKMAKADUR ADGANGUR
Italien	Segretissimo	Segreto	Riservatissimo	Riservato
Kasachstan		Совершенно секретно	Секретно	Для служебного пользования
Lettland	Sevišķi slepeni	Slepeni	Konfidenciāli	Dienesta vajadzībām
Litauen	Visiškai slaptai	Slaptai	Konfidencialiai	Riboto naudojimo
Luxemburg	Très Secret Lux	Secret Lux	Confidentiel Lux	Restreint Lux
Malta	L-Oghla Segretezza Top Secret	Sigriet Secret	Kunfidenzjali Confidential	Ristrett Restricted
Montenegro		TAJNO	POVJERLJIVO	INTERNO
Niederlande	Stg. ZEER GEHEIM	Stg. GEHEIM	Stg. CONFIDENTIEEL	Dep. VERTROUWELIJK
Nord-Makedonien	ДРЖАВНА ТАЈНА	СТРОГО ДОВЕРЛИВО	ДОВЕРЛИВО	ИНТЕРНО



Norwegen	STRENGT HEMMELIG	HEMMELIG	KONFIDENSIELT	BEGRENSET
Österreich	Streng Geheim	Geheim	Vertraulich	Eingeschränkt
Polen	Ściśle Tajne	Tajne	Poufne	Zastrzeżone
Portugal	Muito Secreto	Secreto	Confidencial	Reservado
Rumänien	Strict secret de importanță deosebită	Strict secret	Secret	Secret de serviciu
Schweden	HEMLIG/TOP SECRET HEMLIG oder AV SYNNERLIG BETYDELSE FÖR RIKETS SÄKERHET	HEMLIG/SECRET oder HEMLIG	HEMLIG/CONFIDENTIAL oder HEMLIG	EMLIG/RESTRICTED oder HEMLIG
Schweiz		SECRET oder GEHEIM oder SEGRETO	CONFIEDENTIEL oder VERTRAULICH oder CONFIDENZIALE	INTERNE oder INTERN oder AD USO INTERNO
Slowakei	Prísne tajné	Tajné	Dôverné	Vyhradené
Slowenien	STROGO TAJNO	TAJNO	ZAUPNO	INTERNO
Spanien	SECRETO	RESERVADO	CONFIDENCIAL	DIFUSIÓN LIMITADA
Tschechische Republik	Přísně tajné	Tajné	Důvěrné	Vyhrazené
Ukraine		Цілком таємно	Таємно	
Ungarn	Szigorúan titkos!	Titkos!	Bizalmas!	Korlátozott terjesztésű!



Vereinigte Staaten von Amerika	TOP SECRET	SECRET	CONFIDENTIAL	(siehe Fußnote[3])
Vereinigtes Königreich	UK TOP SECRET	UK SECRET	[4]	UK OFFICIAL-SENSITIVE
Zypern	Άκρως Απόρρητο (Abkürzung: ΑΑΠ)	Απόρρητο (Abkürzung: ΑΠ)	Εμπιστευτικό (Abkürzung: ΕΜ)	Περιορισμένης Χρήσης (Abkürzung: ΠΧ)

Fsas GenAI + MCP for Defense

PrivateGPT Core & Enterprise GenAI Platform

On-premise sovereign AI platform

- Complete data sovereignty
- Air-gapped deployment
- Multi-level security
- RAG technology

+

RAG and/or MCP Integration

Model Context Protocol connectivity

- Connect classified systems
- Legacy system integration
- Real-time data access
- Unified AI interface

Secure, scalable AI for defense operations

© 2025 Fsas Technologies

Key MCP Integration Points

Data Sources

- Air-gapped document repositories
- Classified databases (all levels)
- Legacy systems integration
- Real-time sensor feeds

Operational Systems

- C4ISR platforms
- Equipment/logistics databases
- Training management systems
- Communications systems

Intelligence Systems

- SIGINT/HUMINT databases
- Geospatial/imagery systems
- Threat intelligence platforms
- OSINT aggregators

Security Systems

- Network monitoring tools
- Vulnerability scanners
- Personnel security systems
- Audit and compliance logs

MCP enables unified AI access across heterogeneous classified systems

The Defense AI Advantage

Fsas Technologies
a Fujitsu company

PrivateGPT + MCP

Sovereign AI platform with complete data control

- ✓ Air-gapped security
- ✓ Multi-level classification support
- ✓ Legacy system integration via MCP
- ✓ Real-time operational intelligence
- ✓ Mission-critical reliability

Ready to Deploy

Proven platform with defense-grade security

Scalable Architecture

From tactical units to enterprise deployments

Future-Proof

Continuous model updates and capability expansion

Contact: www.fujitsu.com/emeia/private-gpt

Fsas Technologies

a Fujitsu company

Linkin Park – „Numb“
Chester Bennington
*1976 - †2017