

Private GPT

Data sovereignty and efficiency for enterprises



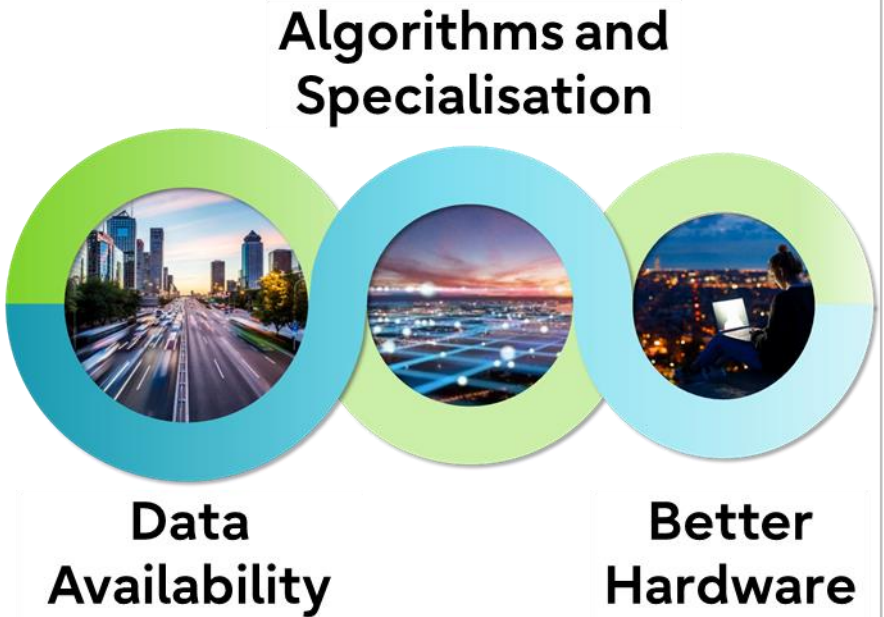
Udo Würtz

CDO Europe
Fellow

AI & Co.

Use Case Families	Generative Models	Non-Generative ML	Optimisation	Simulation	Rules	Graphs
Forecasting						
Planning						
Decision Intelligence						
Autonomous System						
Segmentation						
Recommender						
Perception						
Intelligent Automation						
Anomaly Detection						
Content Generation						
Chatbots						
Knowledge Discovery						

Source: Gartner Data & Analytics Summit Conference 2024



What the private GPT is addressing

Data protection and security

Intellectual property issues

Costs

Control

Bias, fairness and ethics

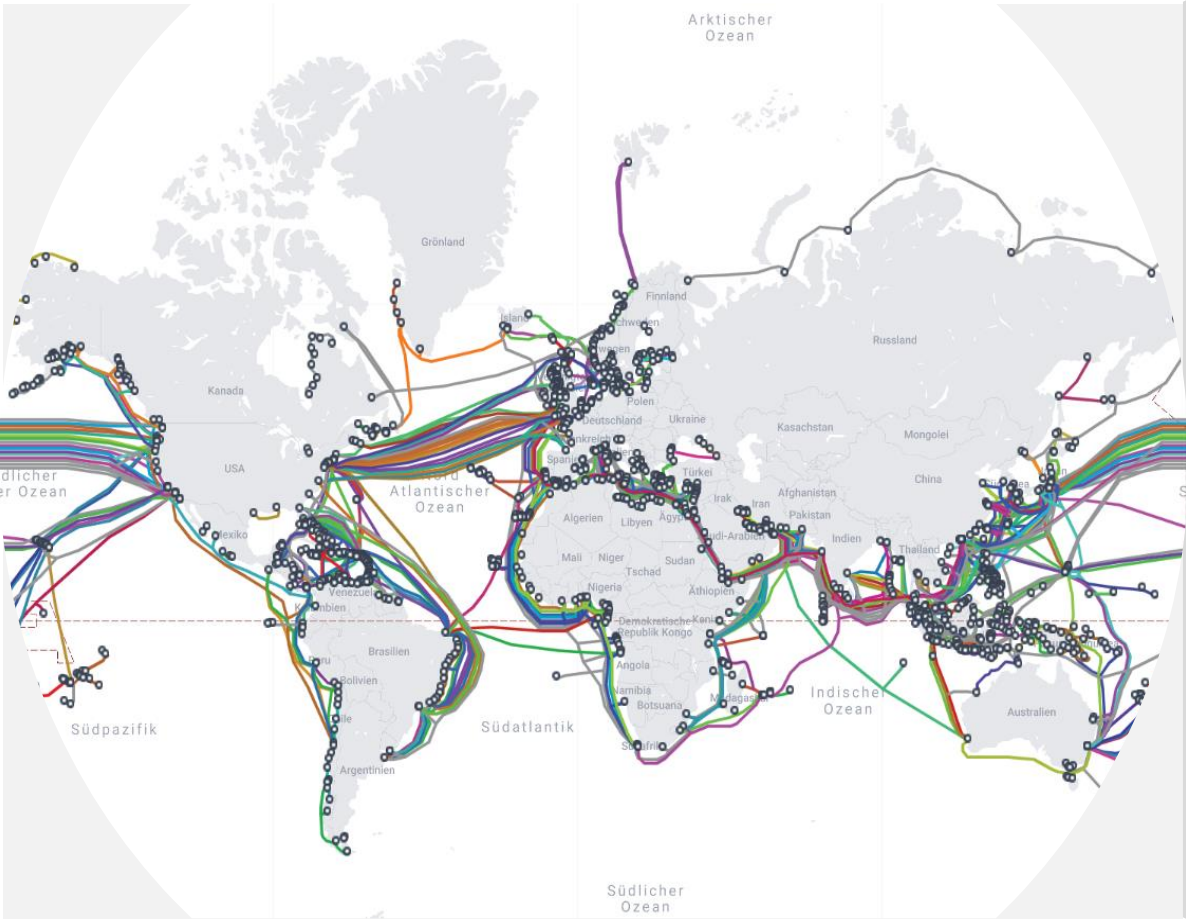
(In)dependence of service providers

Compliance and regulatory challenges

Critical infrastructure



Critical Infrastructure



Kabelsabotage in Frankreich führt international zu Störungen

An mehreren Stellen wurde das Glasfaser-Netz von Free bei Marseille durchtrennt. Die Polizei ermittelt.

[in Pocket speichern](#) [merken](#) [teilen](#)

21. Oktober 2022, 18:55 Uhr, Achim Sawall

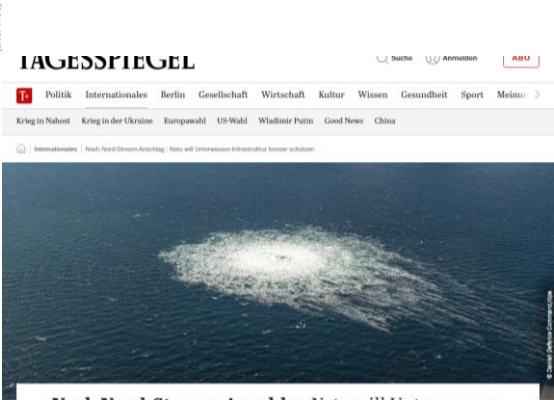


SCHUTZ VOR RUSSISCHER SABOTAGE

"The threat is real"



Die NATO schafft ein neues Zentrum für den Schutz von Gasleitungen und Datenkabeln im Meer. Ein General der Allianz warnt, Russland habe die Infrastruktur schon kartographiert.



Nach Nord-Stream-Anschlag Nato will Unterwasser-Infrastruktur besser schützen

Digitally sovereign workplace

1. Data sovereignty

Full control over data, independent of external providers.

2. Data security

Use of trustworthy, mostly domestic or European solutions that meet strict data protection requirements.

3. Independence

Free choice of technologies and platforms, without dependence on global tech giants.

4. Flexibility

Adaptable workplace that maintains control and sovereignty over digital processes.



Non-Open Source versus Open Source



GLaM (Google)
LaMDA (Google)
Gopher (DeepMind)
Chinchilla (DeepMind)
Ernie 3.0 Titan (Baidu)
HyperCLOVA (Naver Corp)
AlexaTM 20B (Amazon)
Megatron Turing-NLP (Microsoft)
GPT-SW3 (AI Sweden & RISE)
ChatGPT (OpenAI)
GPT 3.5 (OpenAI)
GPT 4 (OpenAI)
GPT 4o (OpenAI)
Gemini (Google)



Wu Dao 2.0 (BAAI)
FLAN (Google)
PaLM (Google)
Bloom (Big Science)
T0-XXL (Big Science)
OPT (Meta)
LLaMA (Meta)
YaLM (Yandex)
GPT-j (EleutherAI)
NeMo (Fujitsu / Mistral)
GPT-NeoX (Eleuther AI)
[Gemma (Google)]

Where are we?

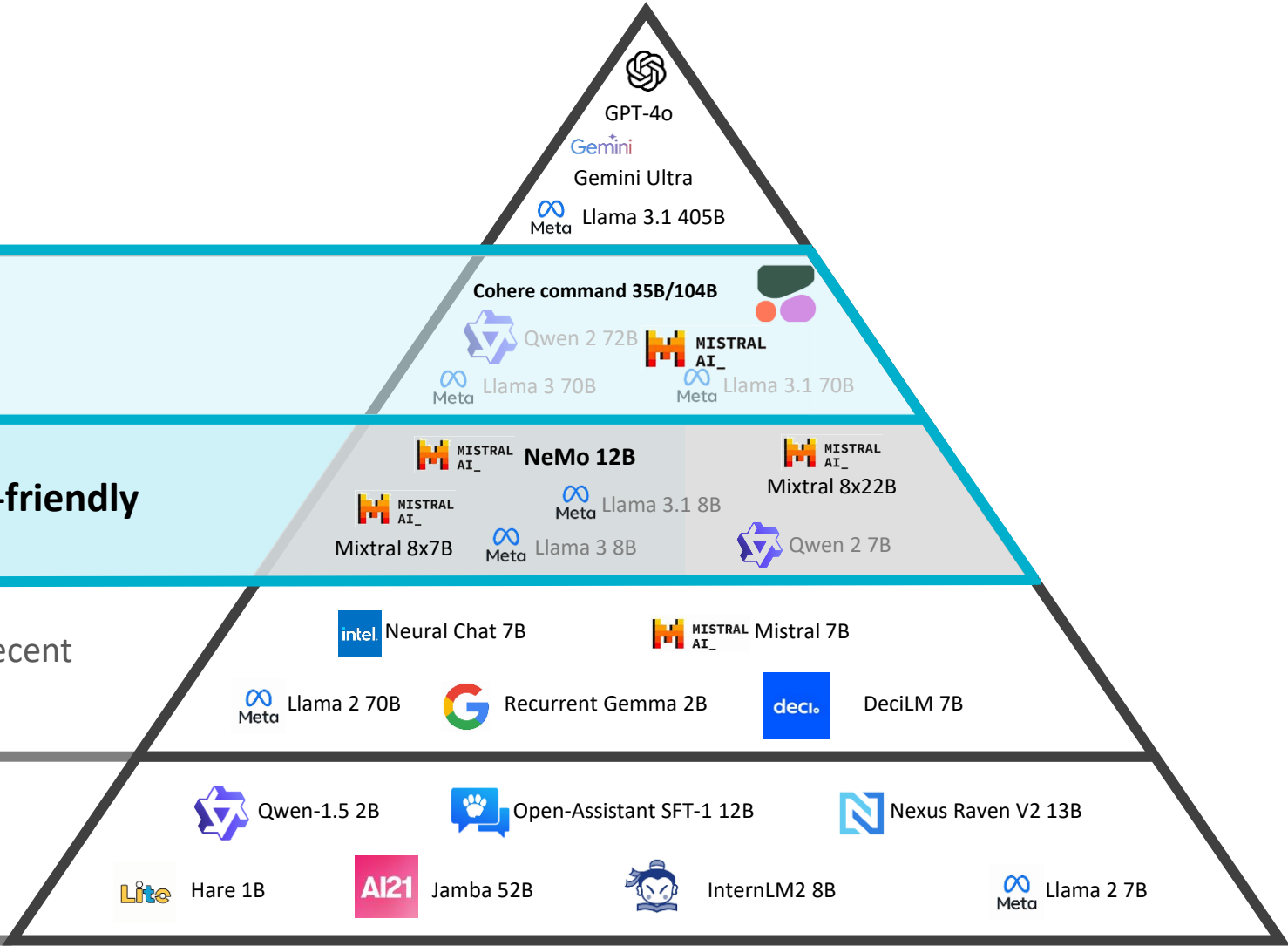
Powerful, very expensive

High-quality, cost-intensive models

Excellent performance, budget-friendly

Good models, performance okay, decent response

Poor performance, which usually doesn't elicit great responses



Quelle: Hugging Face

The maths and mathematicians behind our LLM



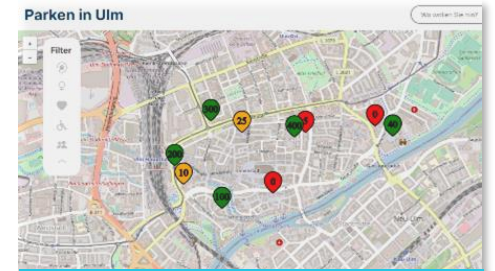
Artificial intelligence



Analytics



Automation



Digitisation



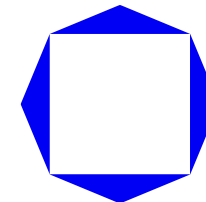
Stadt Ulm



Co-financed by the
European Union



Baden-Württemberg



DASU

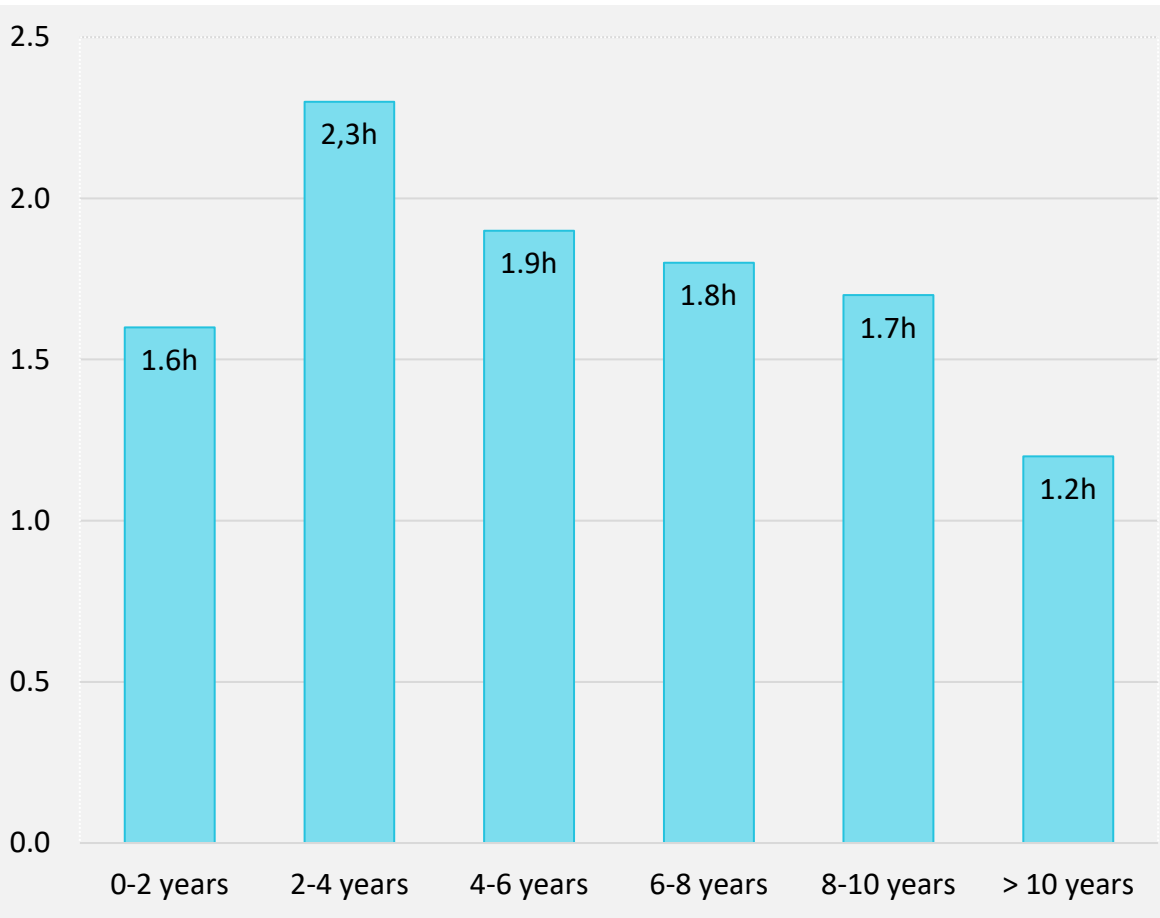
Customers

Use cases and Benefits

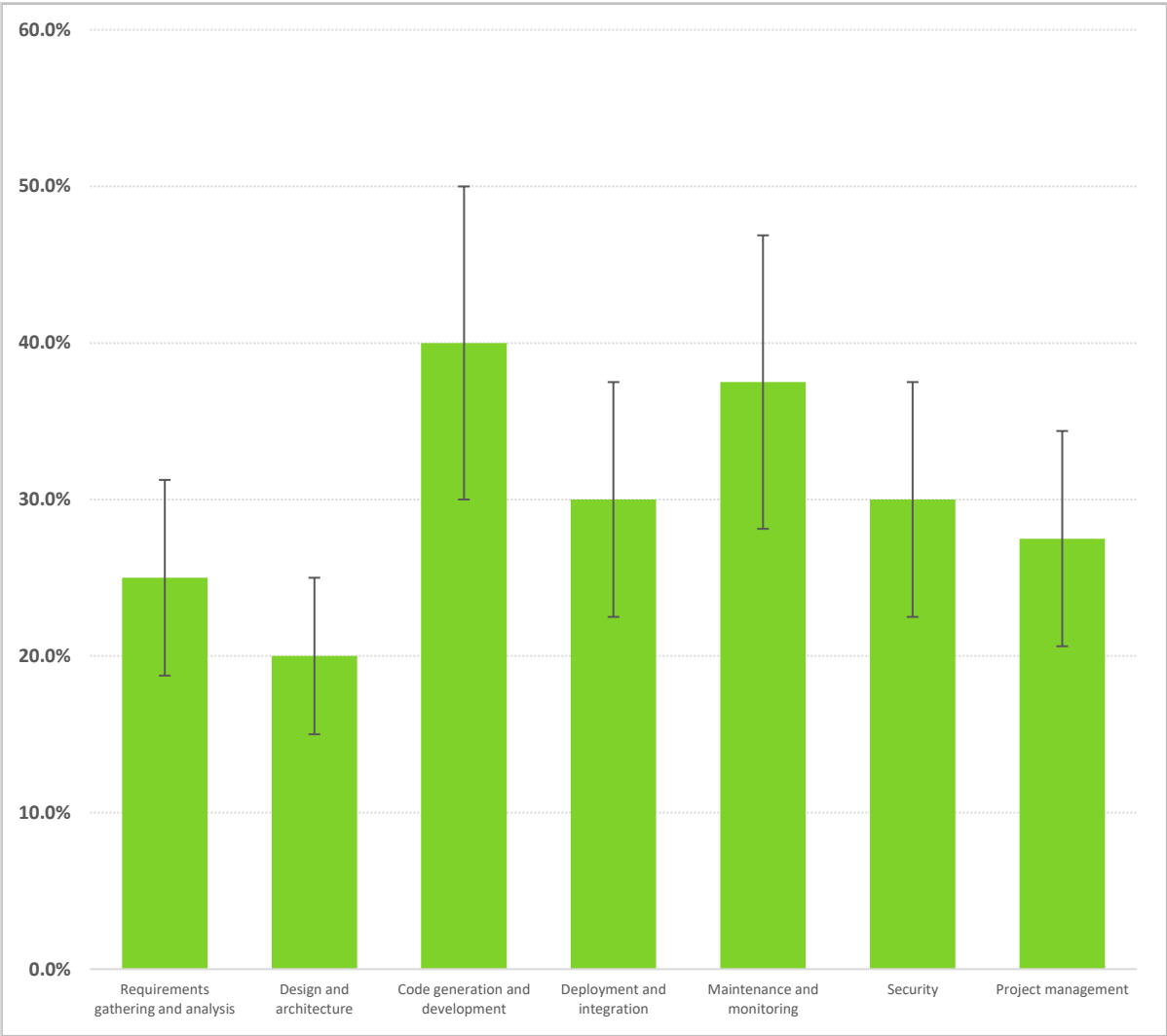


Increased efficiency through generative AI

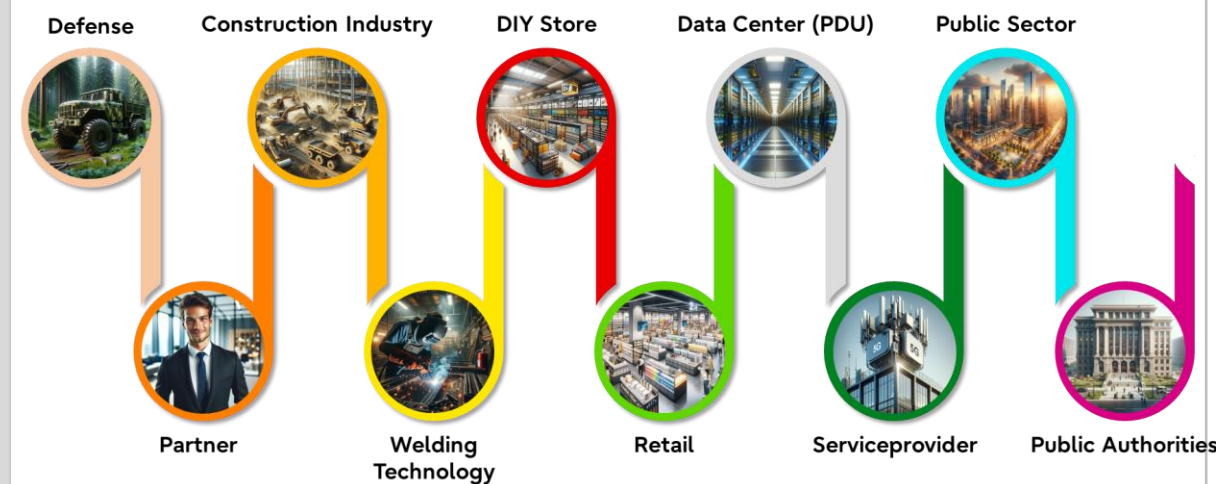
Example: software development at Fujitsu

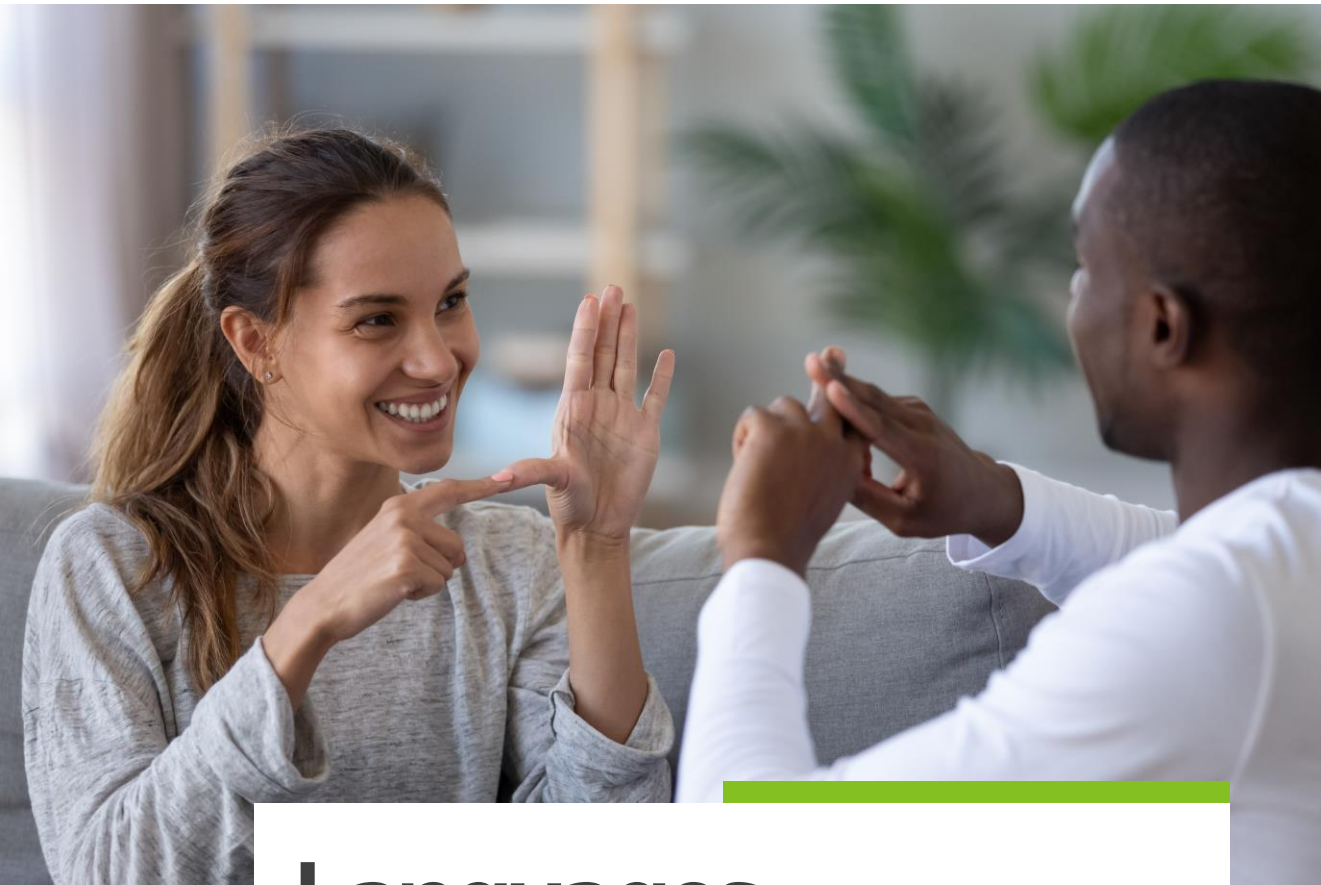


Experience of the developers; efficiency increase in h/day



Customer use cases



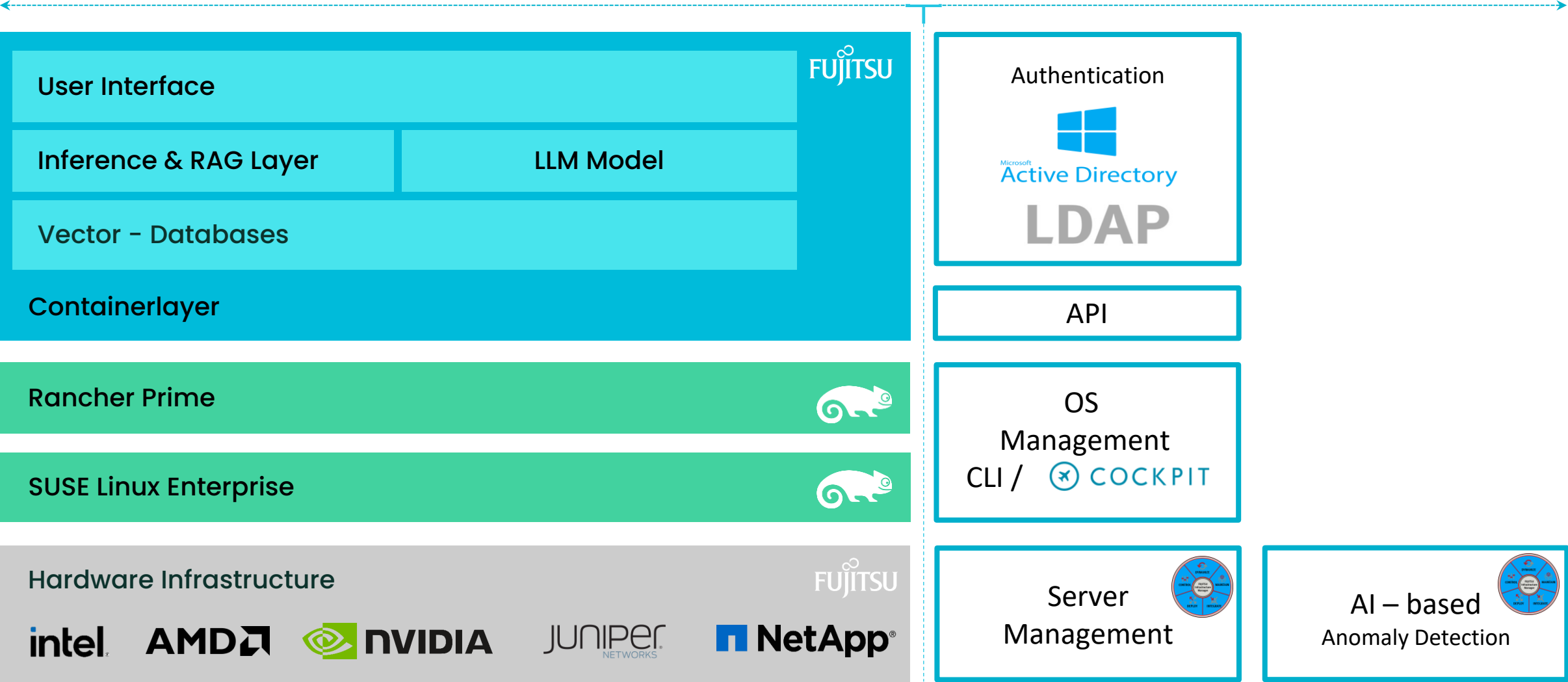


Languages

Natively or Translated

Arabic	Finnish	Japanese	Swedish
Bulgarian	French	Norwegian	Turkish
Chinese	German	Persian	Ukrainian
Czech	Greek	Polish	
Danish	Hebrew	Portuguese	
Dutch	Hindi	Romanian	
English	Hungarian	Russian	
Esperanto	Indonesian	Slovakian	
Estonian	Italian	Spanish	

LLM – Reference Infrastructure



Security threads and CRA Cyber Resilience Act

New EU legislation

NIS2 - Network and Information Security

Directive concerns public and private critical infrastructure facilities (energy, transport, healthcare, water supply, digital infrastructure, public administration, etc.). This has been expanded, e.g. transport, financial market infrastructure, pharma and medtech, chemicals and many more.

Cyber Resilience Act (CRA)

aims to protect consumers and businesses that buy or use products or software with a digital component.

Digital Operational Resilience Act (DORA)

affects all financial and insurance companies and is an EU regulation that came into force on 16 January 2023 and will apply from 17 January 2025.

NIS-2 Directive (in extracts)

- Up to €10 million liability or 2% of turnover
- Management can be held personally liable

NIAP OSPP (US) vs. EAL 4+ (EU) - Scope of certification

EAL = Evaluation Assurance Level (how much was checked)

Security certification gap for NIS2

NIAP OSPP
on level EAL 1



Product

US-Companies, SUSE

EAL 4+

Security updates

Production

Company

Product

SUSE

Supply chain

Features

Impact of GDPR and AI Act on e.g. Meta's LLaMA 3 Release

Restrictions of LLaMA 3 in the EU

- **The Reason:**
The EU AI Act strictly regulates high-risk AI systems.
- **Metas Decision:**
Meta will not release LLaMA 3 in the EU due to GDPR and AI Act.
- **Consequences:**
EU companies are barred from using Meta's AI models.
No-EU companies are restricted to sell to EU customers



Data chat

API and more



PDFs



Filter




Search ...



Showing 71 / 71



Upload PDF


 Annual_Report_NISD_Security_Incidents_2020_ZBjUJx25UOZIDesO...

24.10.2024 10:49

1.19 MB

EU_AI_NIS_Incidents ×




 Annual_Report_NISD_Security_Incidents_2021_fQTKZ82dzWRUTU1I...

24.10.2024 10:49

1.34 MB

EU_AI_NIS_Incidents ×



 Annual_Report_NISD_Security_Incidents_2022_Mq6UxAADvOwkGx...

24.10.2024 10:49

1.68 MB

EU_AI_NIS_Incidents ×



 C_2023_6068_F1_COMMUNICATION_FROM_COMMISSION_EN_V...

24.10.2024 10:44

174.18 KB

EU_NIS_2 ×




 C_2023_6070_F1_COMMUNICATION_FROM_COMMISSION_EN_V...

24.10.2024 10:45

190.74 KB

EU_NIS_2 ×



 CELEX_02022L2555-20221227_EN_TXT.pdf

24.10.2024 10:28

655.59 KB

EU_NIS_2 ×




 CELEX_32022L2555_EN_TXT.pdf

24.10.2024 10:19

1.26 MB

EU_NIS_2 ×




 CELEX_32022R2554_EN_TXT.pdf

24.10.2024 11:33

1.42 MB

EU_DORA ×




 cg_publication_03_20_-_annual_report_nisd_security_06A30C2C-DA...

24.10.2024 10:49

999.36 KB

EU_AI_NIS_Incidents ×



 cg_publication_03_20_-_annual_report_nisd_security_06A30C2C-DA...

24.10.2024 10:49

999.36 KB

EU_AI_NIS_Incidents ×



+ New Chat

Search ...

October 2024

- please summarize nis 2
- i am selling cables, will i be aff...
- What is NIS 2 meaning?

Select your knowledge base

Group documents

General knowledge

Group documents

✓ Group knowledge

× No follow-up questions

× No general knowledge

When enabled, the AI will reference documents shared with your group to provide more relevant answers in the chat. These documents are managed by the admin and tailored to your group's needs.

Select documents from one or more groups

Public documents

EU_AI_ACT ✓

EU_AI_NIS_Incidents ✓

EU_DORA ✓

EU_NIS_2 ✓

Language

Slovenčina

Native Language

When selected, your prompt is processed directly in the original language without translation. The AI uses models specifically trained for that language, providing responses naturally without any additional translation steps. Please note that the quality of these models can vary.

Translator

When enabled, your prompt is translated into English before being sent to the AI. The AI processes your query in English and then translates the response back to your original language.

Enter text...

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

[+ New Chat](#)

October 2024

please summarize nis 2



i am selling cables, will i be aff...

What is NIS 2 meaning?

Chat language: English

Documents from these groups: [Show all](#)

please summarize nis 2

24.10.2024 10:37



[Copy answer](#)

The NIS 2 Directive (EU) 2022/2555, adopted on 14 December 2022, aims to establish a high common level of cybersecurity across the Union to improve the functioning of the internal market. It introduces measures such as:

- National Cybersecurity Strategies:** Member States must adopt national cybersecurity strategies.
- Competent Authorities:** Member States must designate or establish competent authorities, cyber crisis management authorities, single points of contact on cybersecurity, and computer security incident response teams (CSIRTs).
- Risk Management and Reporting:** Entities of a type referred to in Annex I or II, as well as critical entities under Directive (EU) 2022/2557, must implement cybersecurity risk management measures and report incidents.
- Information Sharing:** Rules and obligations on cybersecurity information sharing are established.
- Supervision and Enforcement:** Member States have supervisory and enforcement obligations.

The directive applies to medium and large entities of types listed in Annex I or II that provide services or carry out activities within the Union.

Sources:

[EU_NIS_2 - CELEX_02022L2555-20221227_EN_TXT.pdf, Page 1](#)


[EU_NIS_2 - CELEX_02022L2555-20221227_EN_TXT.pdf, Page 2](#)


please summarize nis 2 and the impact to companies in detail

[+ New Chat](#)

October 2024



 i am selling cables, will i be aff... 

 please summarize nis 2

 i am selling cables, will i be affected...


 What is NIS 2 meaning?

 Chat language: English

 Documents from these groups: [Show all](#) 

i am selling cables, will i be affected by DORA?


29.10.2024 22:01


 Copy answer

No, you will not be affected by DORA (Regulation (EU) 2022/2554) as it is specifically targeted at financial entities and critical ICT third-party service providers (CTPPs) that provide services to these entities. Since you are selling cables, which is not a service provided to financial entities, DORA does not apply to your business.

Sources:


EU_NIS_2 - C_2023_6068_F1_COMMUNICATION_FROM_COMMISSION_EN_V2_P1_2889869_RMQvEzHZxas8eoJpPF3BpudqE_98286.PDF, Page 11 

EU_DORA - PI_COM_C(2024)902_EN_TXT.pdf, Page 2 

EU_DORA - PI_COM_C(2024)896_EN_TXT.pdf, Page 2 

I am providing loans to customers, will I be affected by DORA?


29.10.2024 22:01


 Copy answer

Based on the context provided, if you are a credit institution providing loans to customers, then yes, you will be affected by the Digital Operational Resilience Act (DORA). This is because credit institutions are among the financial entities covered by DORA, as mentioned in point 2 of the context. Therefore, you will need to comply with the provisions of DORA relating to ICT risk management, incident reporting, digital operational resilience testing, information-sharing arrangements, and ICT third-party risk, among others.

Sources:

EU_NIS_2 - C_2023_6068_F1_COMMUNICATION_FROM_COMMISSION_EN_V2_P1_2889869_RMQvEzHZxas8eoJpPF3BpudqE_98286.PDF, Page 11 

EU_DORA - PI_COM_C(2024)902_EN_TXT.pdf, Page 2 

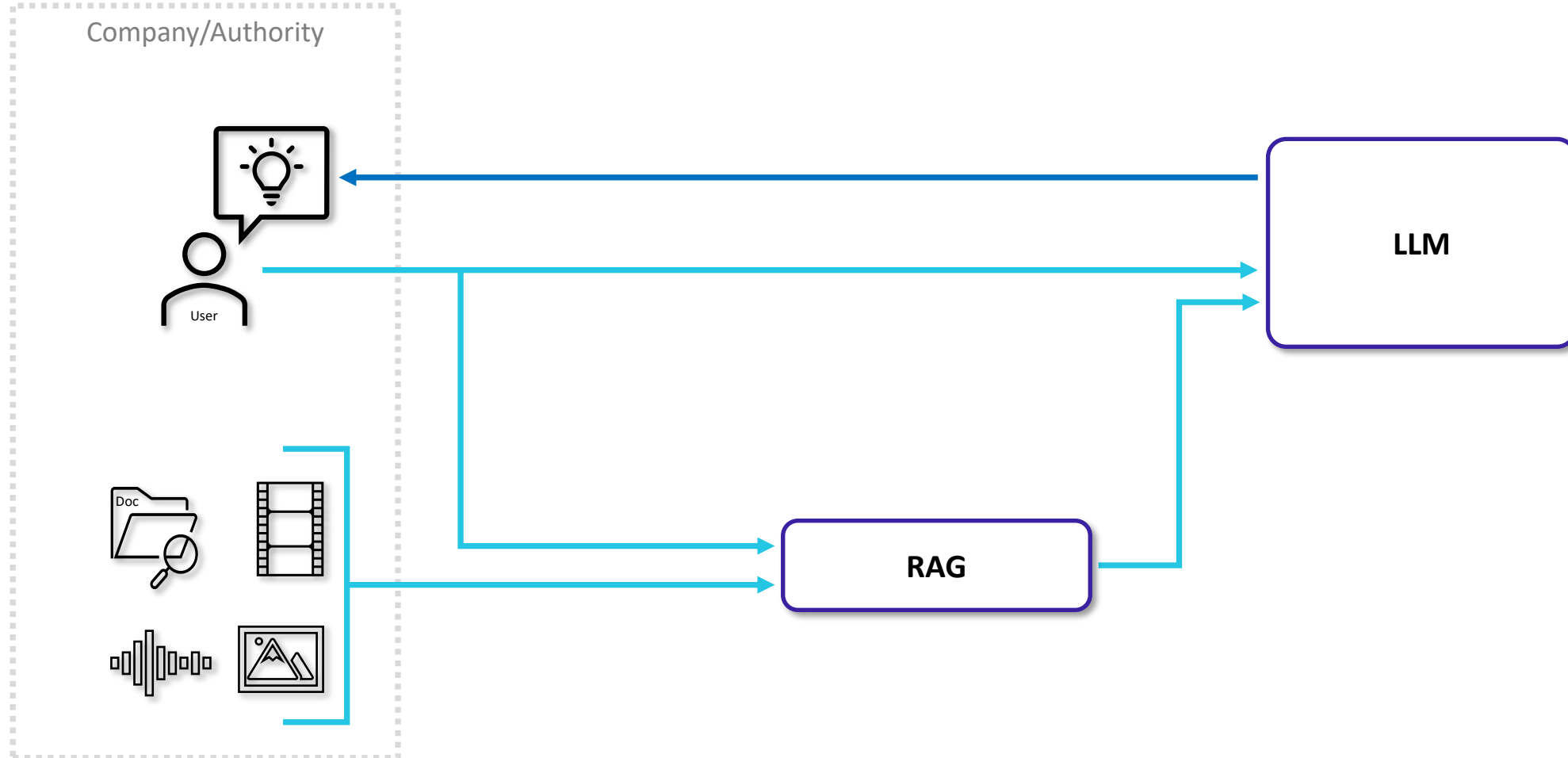
EU_DORA - PI_COM_C(2024)896_EN_TXT.pdf, Page 2 

Enter text...

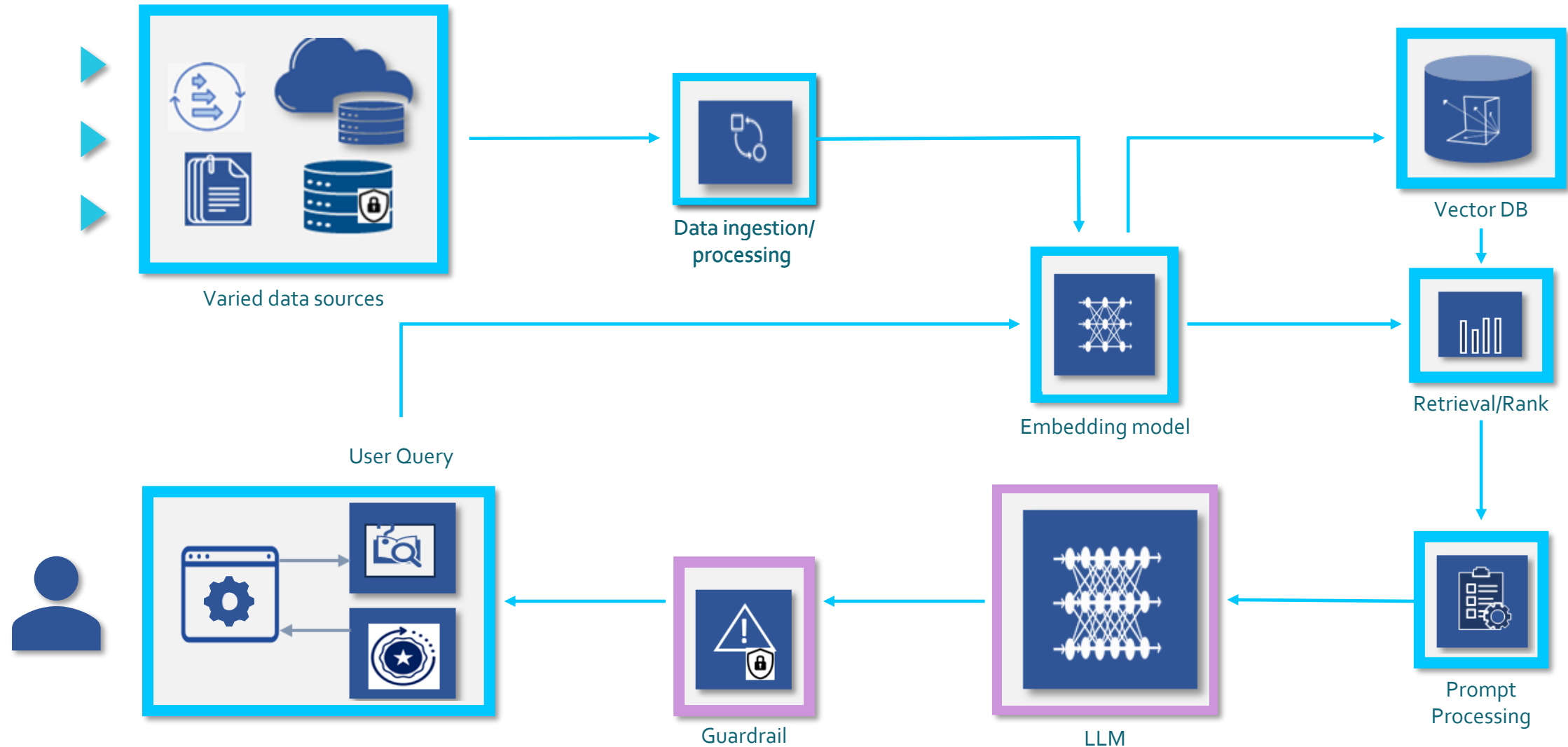


How does our Retrieval Augmented Generation (RAG) work?

Overview

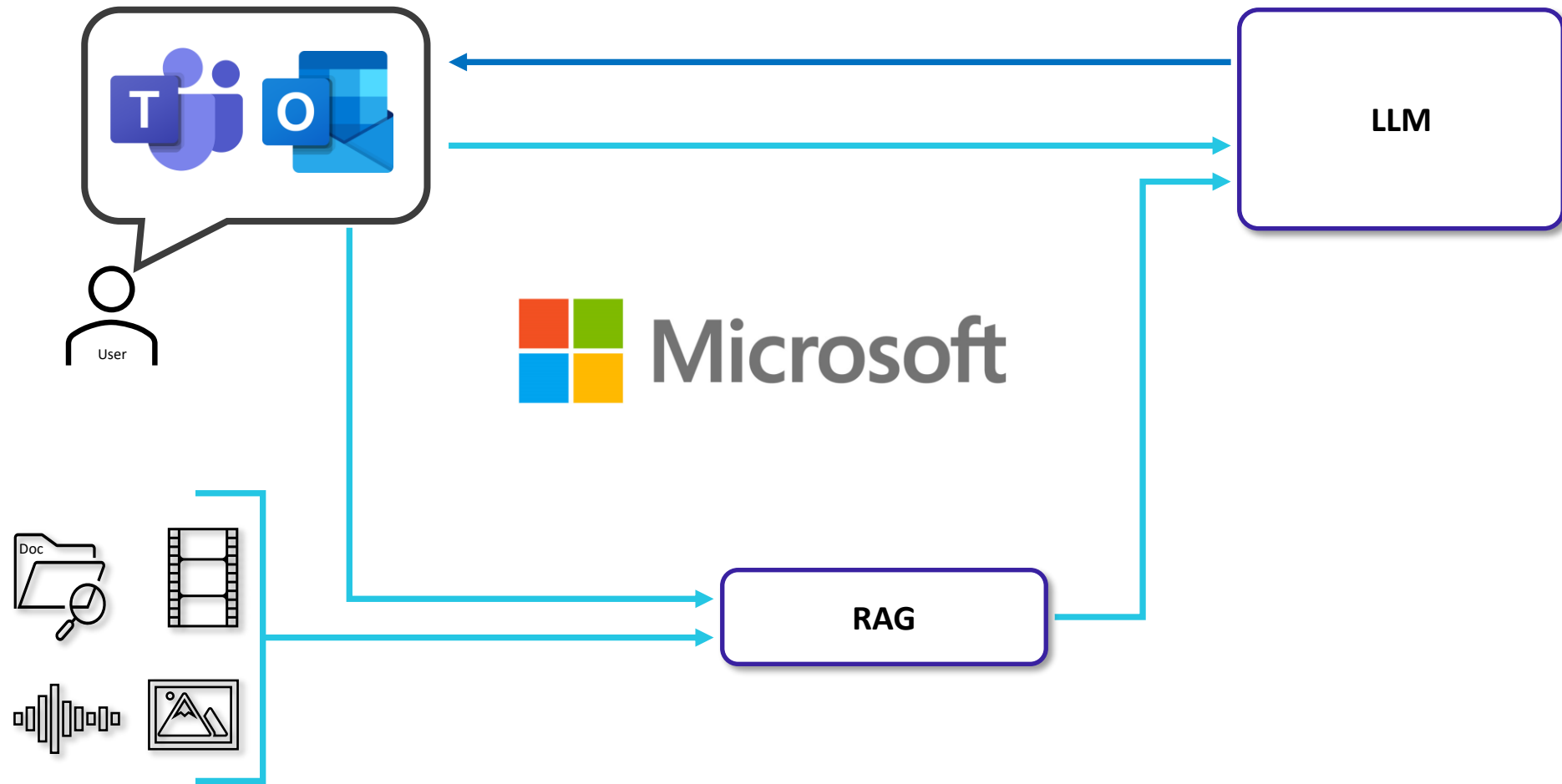


How does Retrieval Augmented Generation (RAG) work?



What customers want

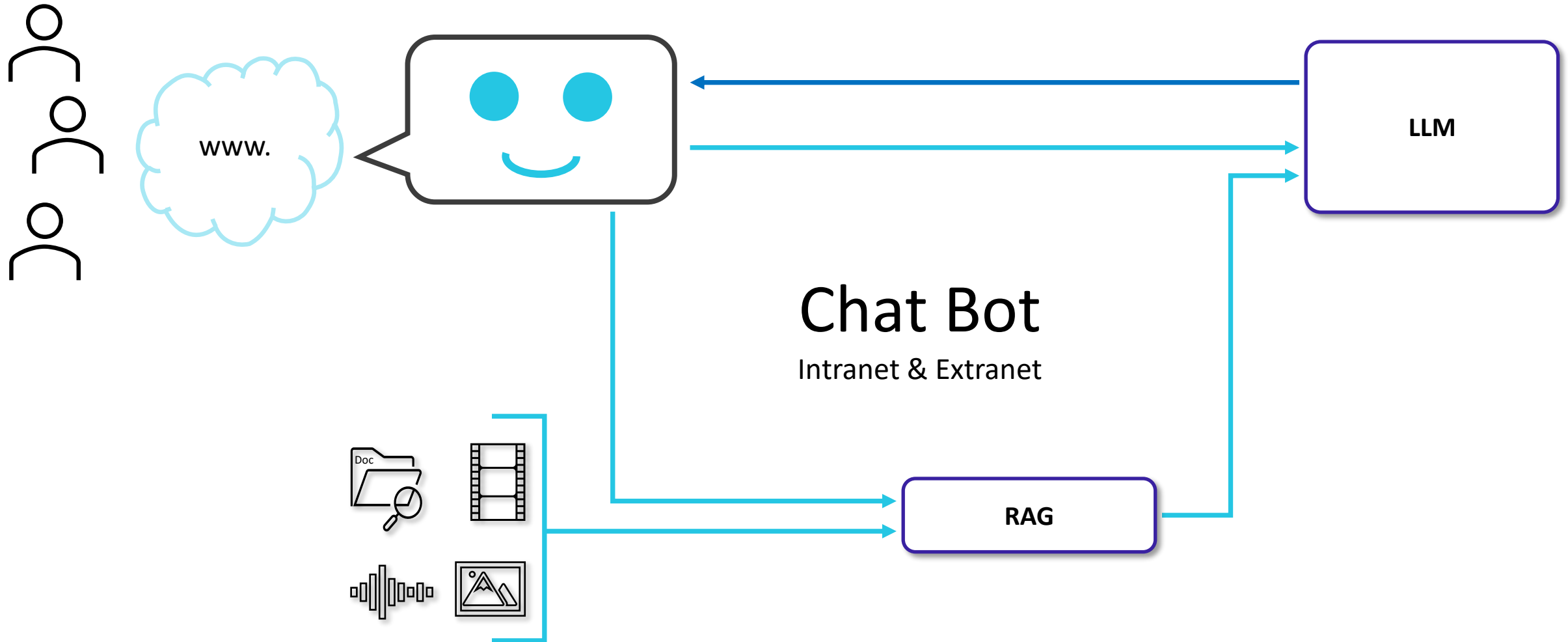
Office Integration



Der Entwurf wurde um 10:48 gespeichert

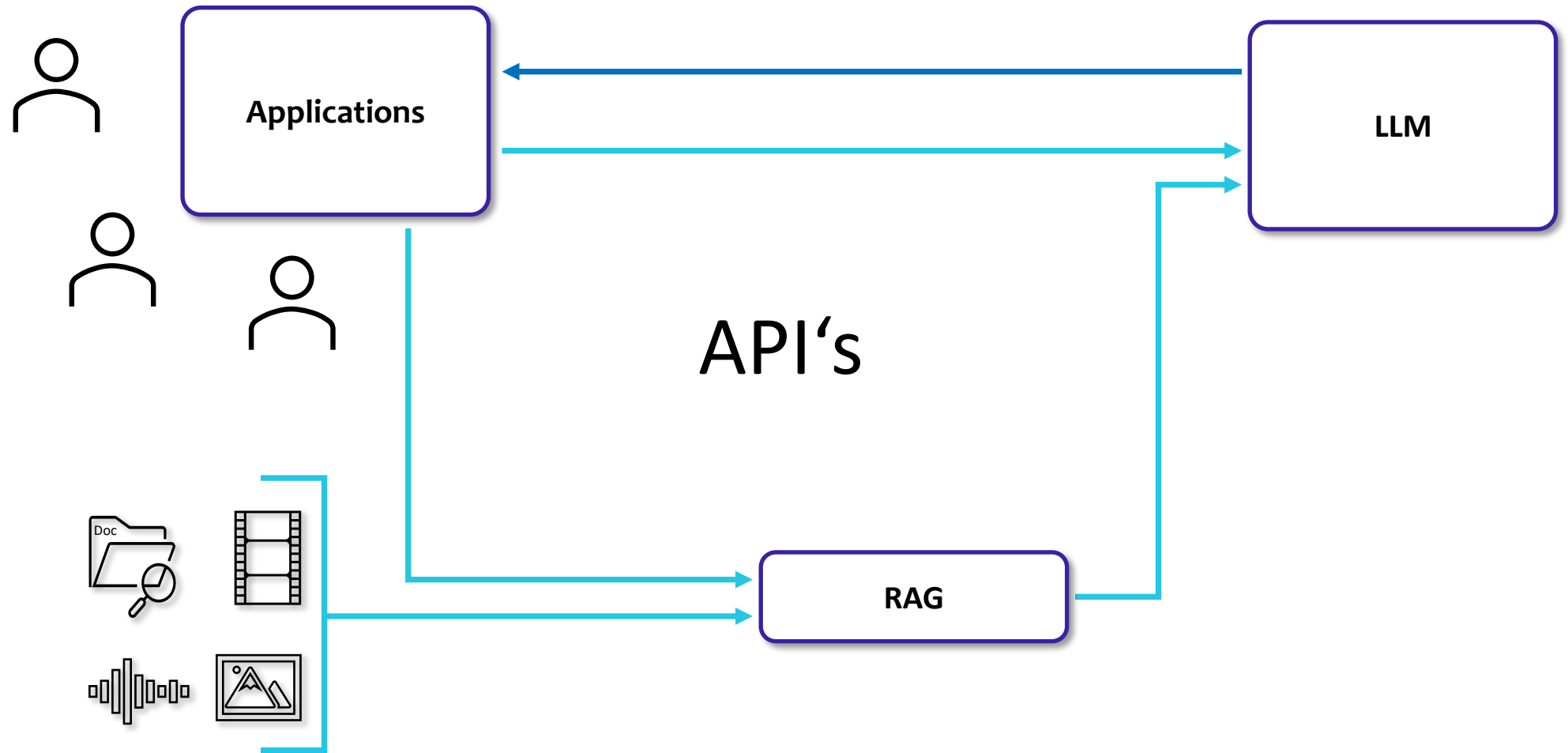
What customers want

Chat Bot Integration



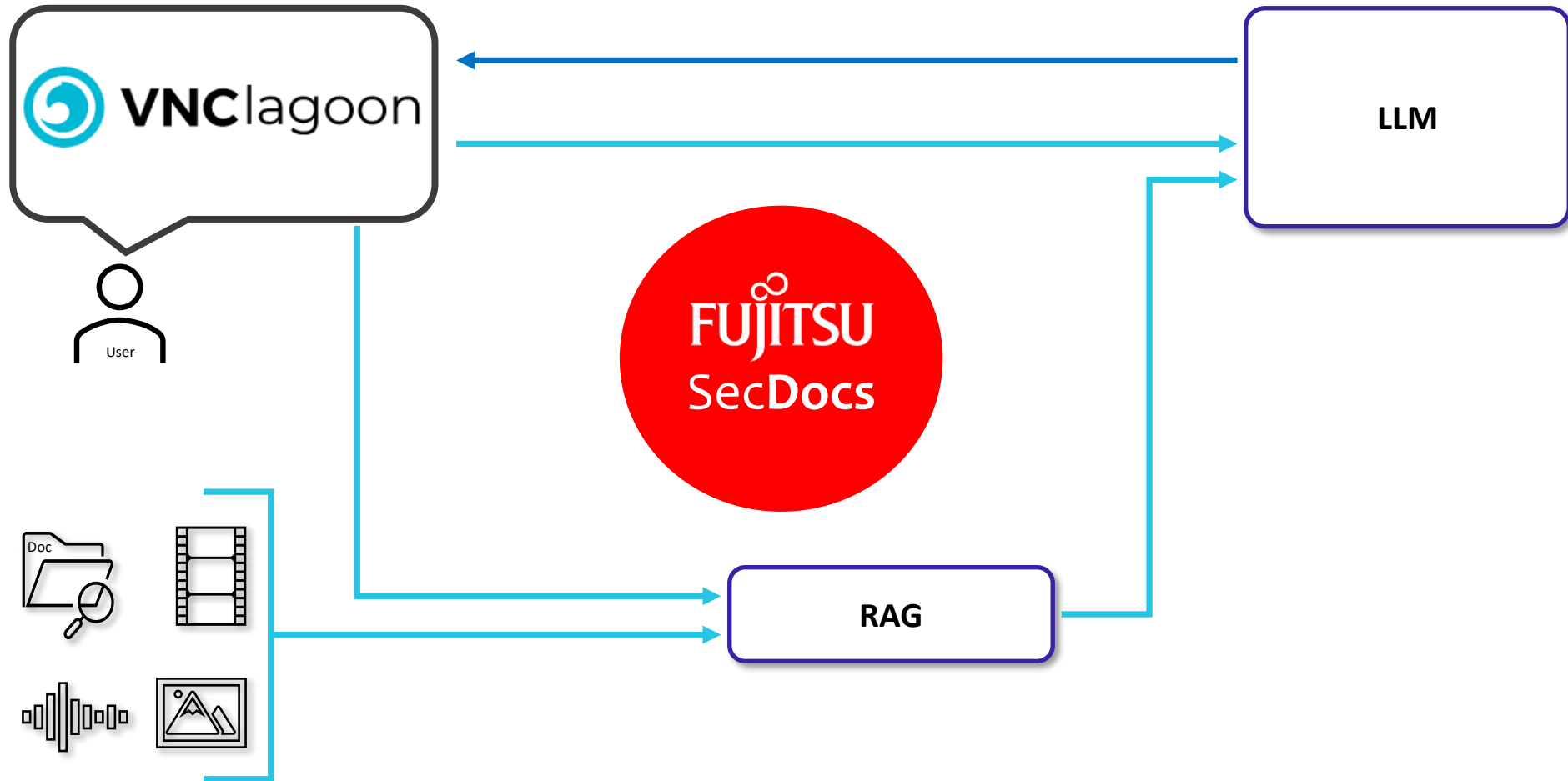
What customers want

APIs facilitate integration.

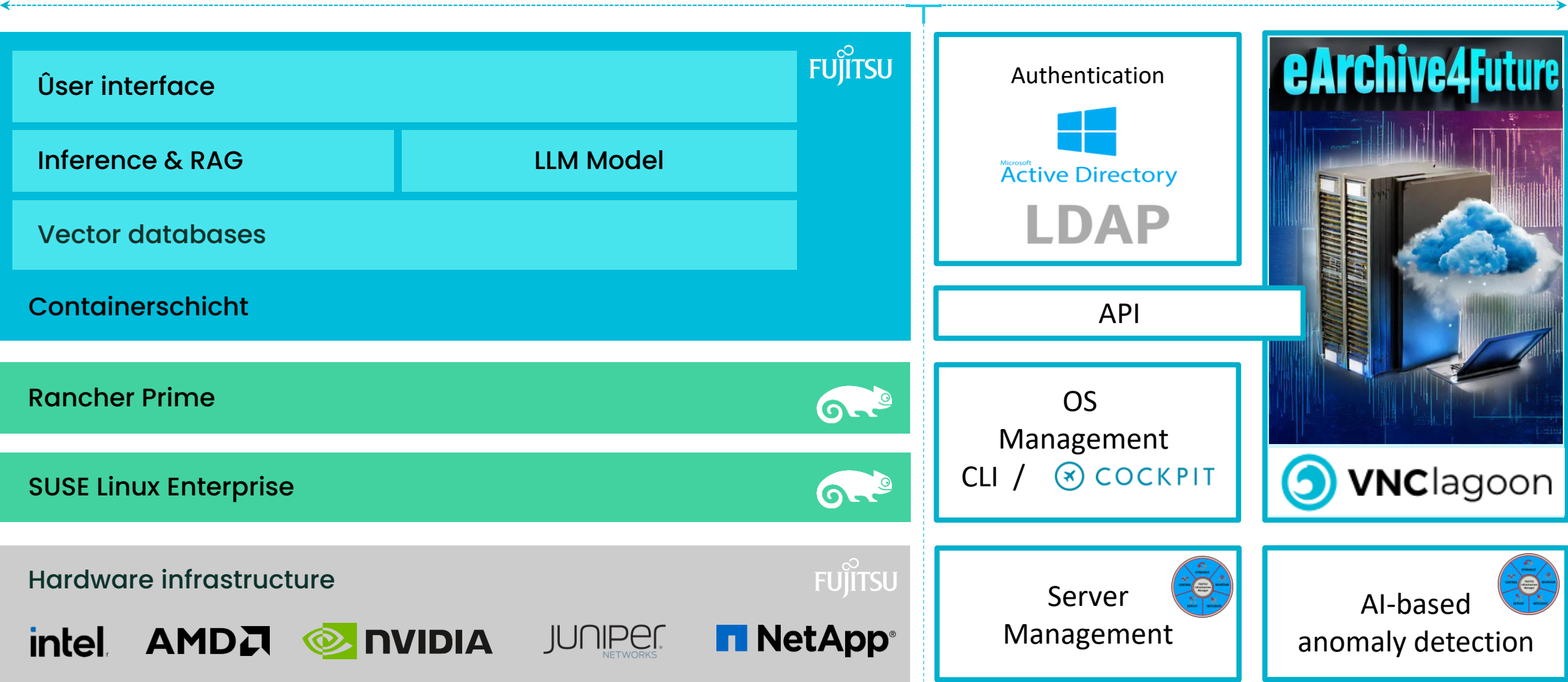


What customers want

SecDocs Integration via VNClagoon



LLM – Reference Infrastructure





Insights

Market & Technology



Mistral NeMo 12B



&



NVIDIA

New model

Mistral NeMo in collaboration with NVIDIA, trained on 3,072 H100 80GB GPUs.

Easy to implement

Can replace any Mistral model.

Multilingual

Strong language skills for multilingual tasks.

Optimised

High performance on NVIDIA hardware and software.

Selected data

Trained on Mistral's own data set, which contains a large proportion of multilingual and coded data. This enables better feature learning, lower bias and an improved ability to handle diverse and complex scenarios.

Data. Bias. Copyright.

The data used to train the model is important.

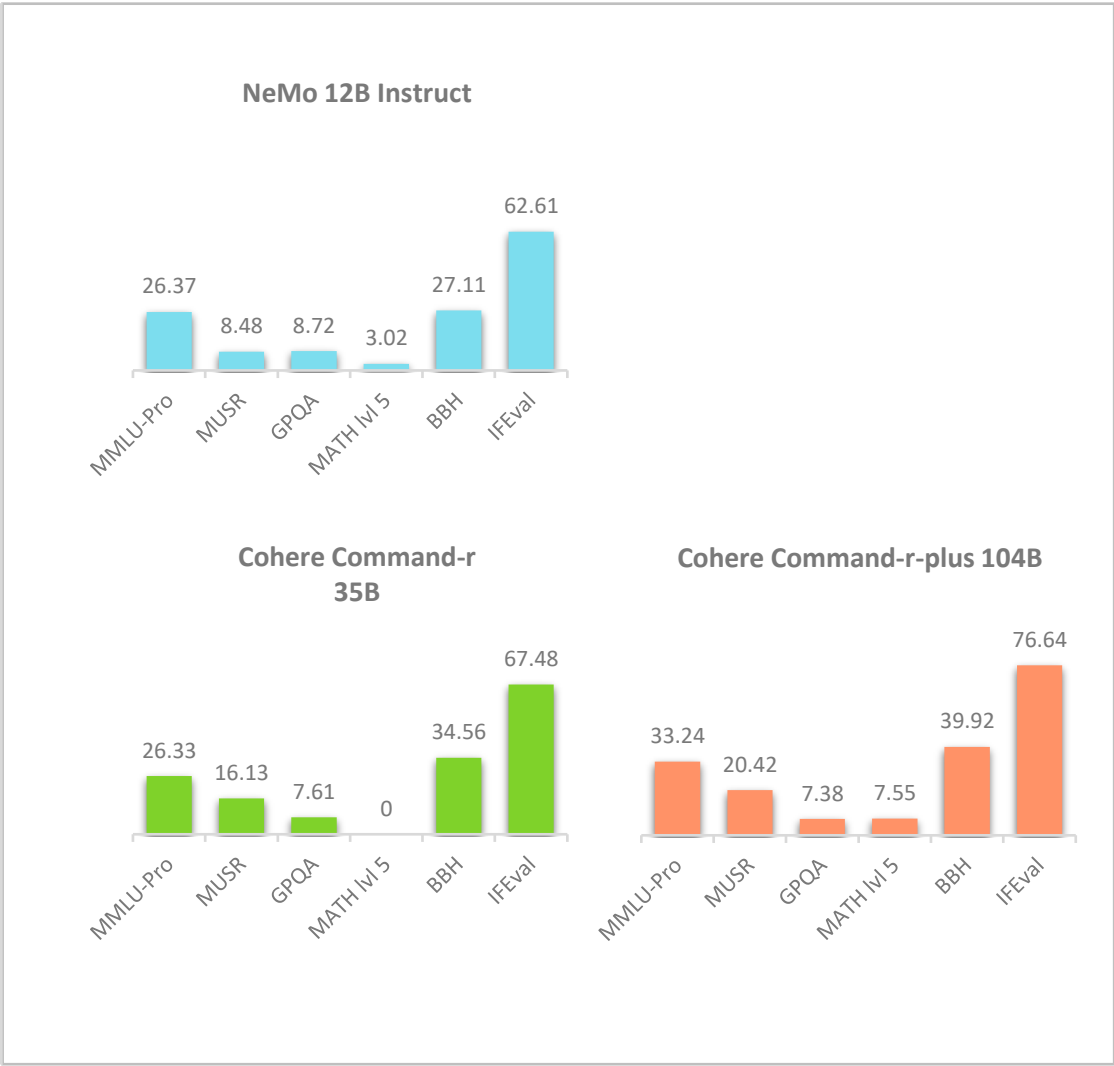
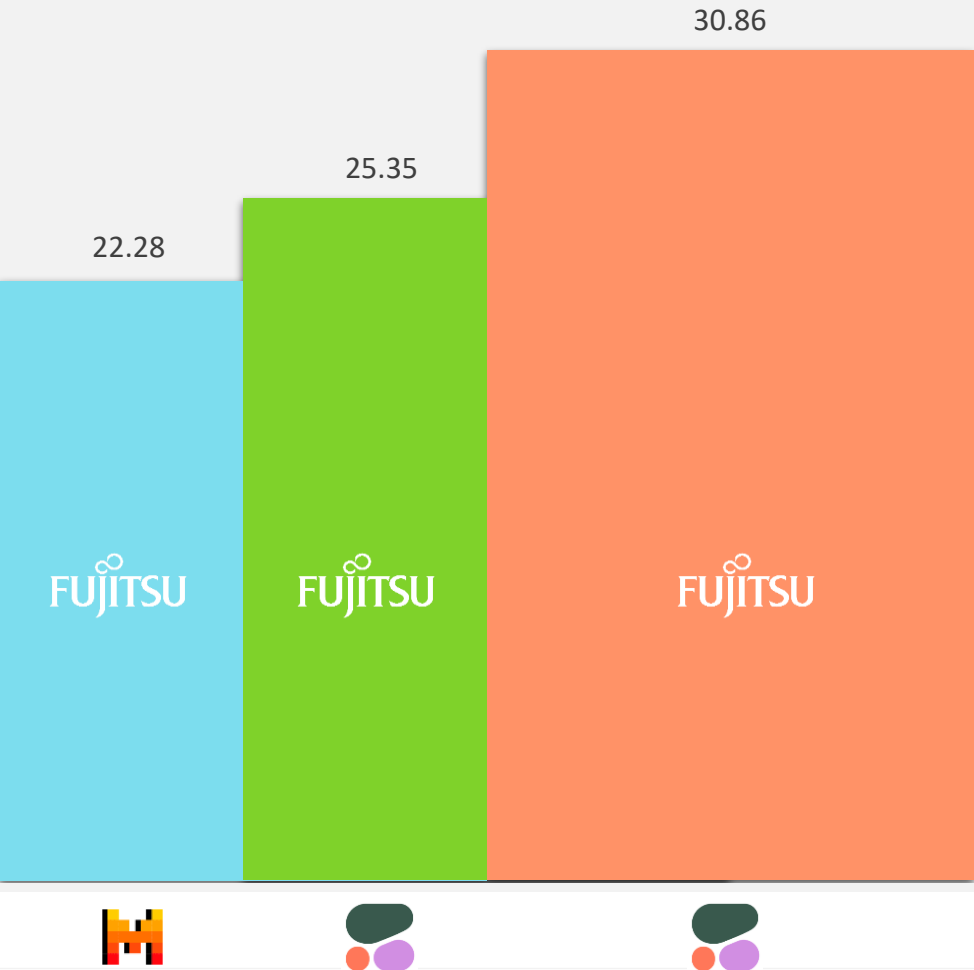


‘Trained on Mistral's proprietary dataset, which contains a large proportion of multilingual and coded data, enabling better feature learning, lower bias and improved ability to handle diverse and complex scenarios.

Typical meaning of proprietary:

...if a data set is labelled as ‘proprietary’, this means that the company, in this case Mistral, has control over the creation, collection, maintenance and management of the data. The company generally ensures that the data comes from legal and licensed sources and guarantees that the use of this data complies with the legal framework...

Our portfolio: a wide selection of high-performance models



Benchmarks

How we compare



Comparison of current models



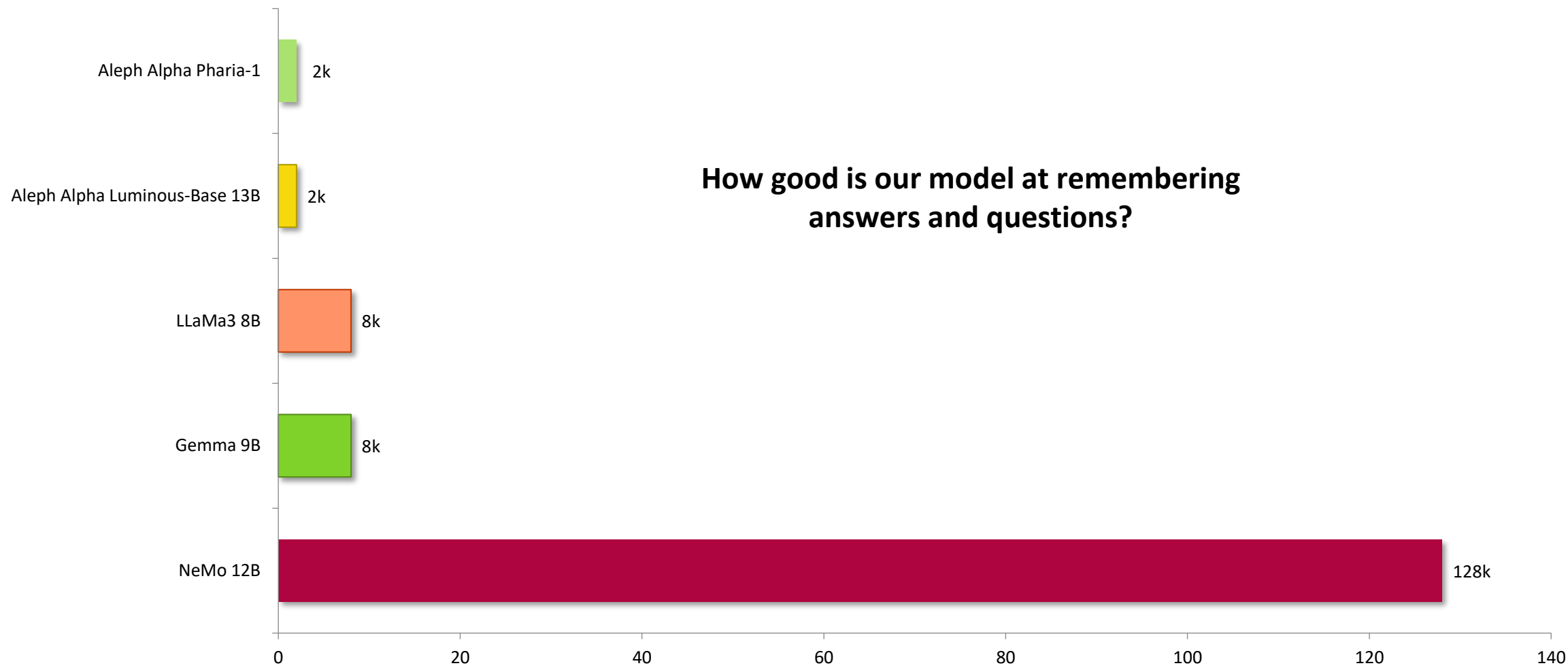
e.g. Huggingface LLM Leaderboard v1

	GPT-3.5	Mixtral 8x22B Instruct	Mixtral 8x7B	Mistral 7B Instruct 0.3	Pharia-1 7B control-aligned	Pharia-1 7B control
MMLU	70,0%	77,8%	70,6%	62,4%	52,5%	48,4%
HellaSwag	85,5%	89,1%	86,7%	82,6%	76,1%	64,6%
ARC Challenge	85,2%	72,7%	85,8%	61,3%	52,8%	54,6%
WinoGrande	81,6%	85,2%	81,2%	78,4%	64,3%	65,1%
GSM-8K	57,1%	58,4%	58,4%	48,8%	16,3%	1,4%
TruthfulQA		84,1%		78,4%	56,6%	54,7%
AVG	75,9%	76,5%	76,5%	68,6%	53,1%	48,1%

Source:
Handelsblatt
Mistral [3]

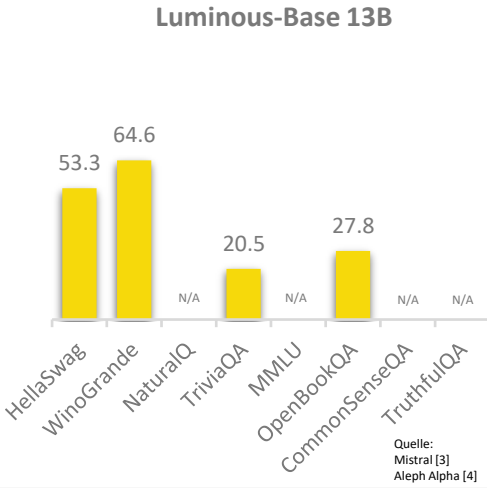
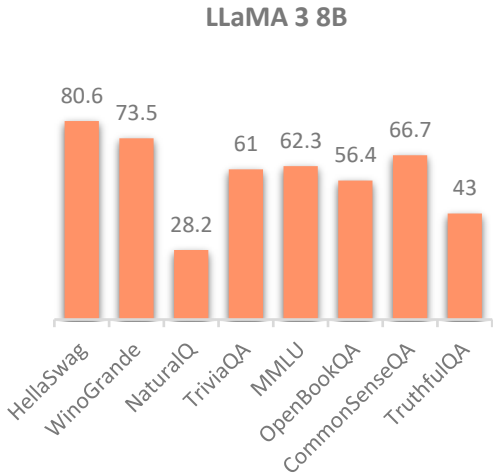
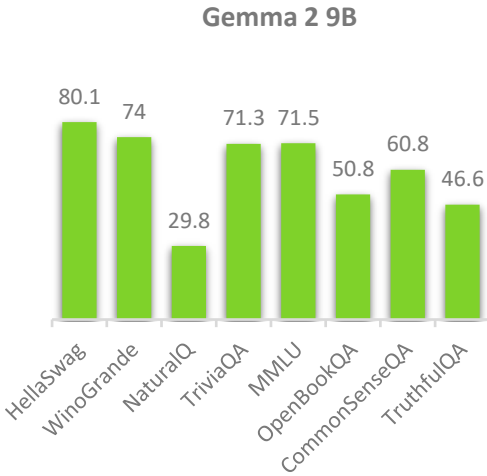
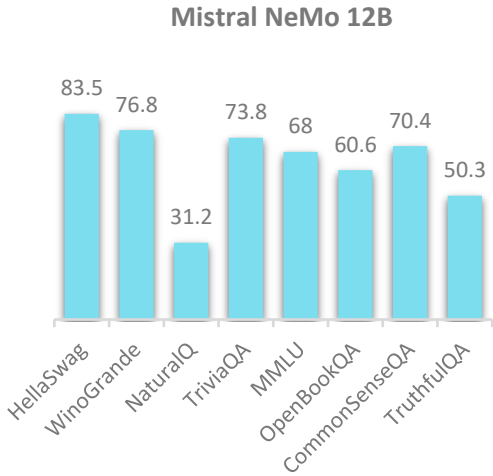
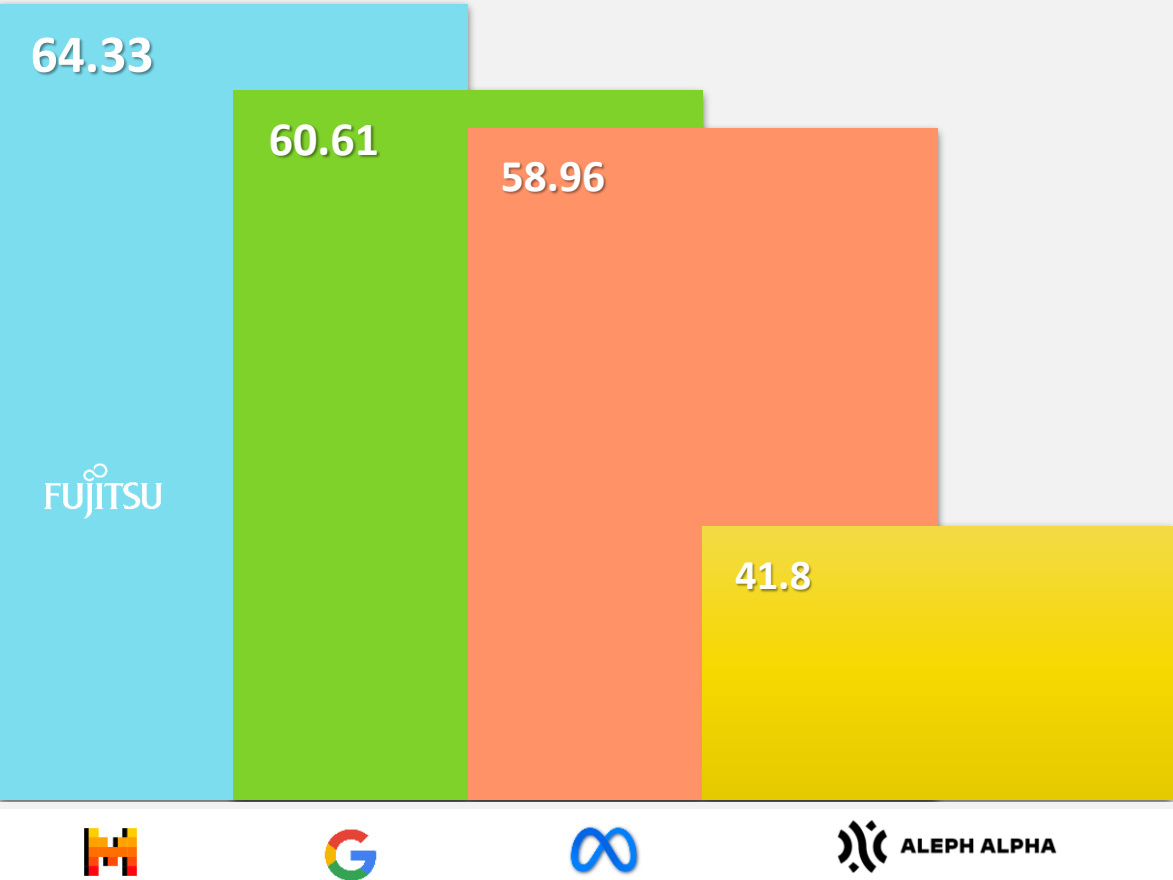
Context Windows

Maximum number of tokens a language model can process at once; allowing a natural conversation.



Source: Mistral [3]

Benchmark of relevant 9B – 13B models



DX Innovation Platform Germany

Your AI Test Drive platform

- No Cloud
- Free of charge
- Consultancy-led
- Secure and reliable
- Available within the EU



intel

AMD

SUSE

RANCHER
PRIME

NVIDIA

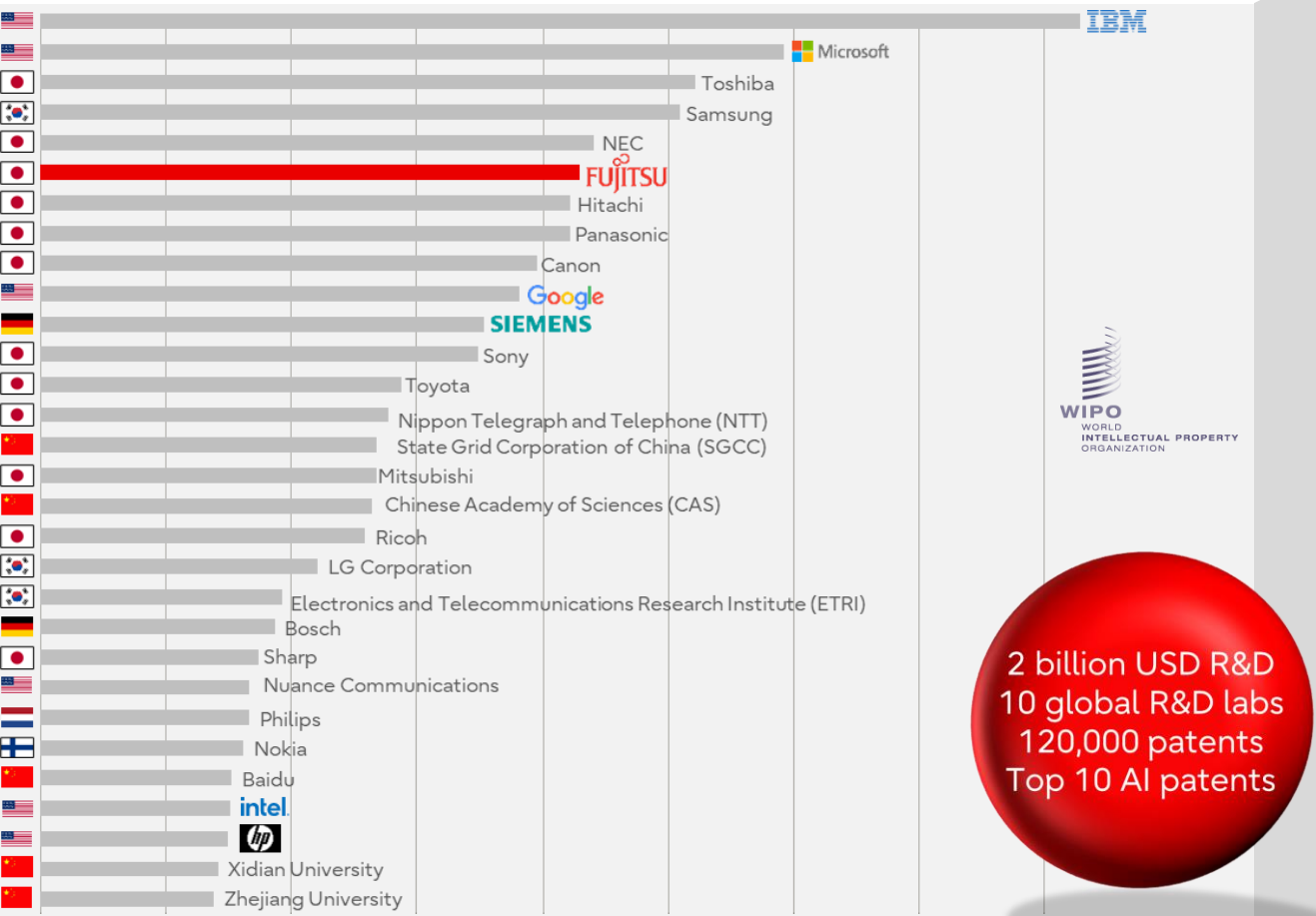
JUNIPER
NETWORKS

NetApp

FUJITSU

Our Ranking

Publication by the WIPO



IBM



2 billion USD R&D
10 global R&D labs
120,000 patents
Top 10 AI patents

Table 9 Top 25 S&T clusters by S&T intensity, 2023

Rank per-capita ^a	Cluster name	Economy	Top Applicant	Top scientific organization
1	Cambridge	GB	ARM	Cambridge University
2	San Jose–San Francisco, CA	US	Google	Stanford University
3	Oxford	GB	Oxford University	Oxford University
4	Eindhoven	NL	Philips Electronics	Eindhoven University of Tech.
5	Boston–Cambridge, MA	US	MIT	MIT
6	Daejeon	KR	LG Chem	KAIST
7	Ann Arbor, MI	US	University of Michigan	University of Michigan
8	San Diego, CA	US	Qualcomm	University of California San Diego
9	Seattle, WA	US	Microsoft	University of Washington Seattle
10	Munich	DE	BMW	Technical University of Munich
11	Kanazawa	JP	Fujitsu	Kanazawa University
12	Raleigh, NC	US	Duke University	Duke University
13	Göteborg	SE	LM Ericsson	University of Gothenburg
14	Beijing	CN	BOE Technology	Tsinghua University
15	Stockholm	SE	LM Ericsson	Karolinska Institutet
16	Helsinki	FI	Nokia	University of Helsinki
17	Zürich	CH	ETH Zürich	ETH Zürich
18	Tokyo–Yokohama	JP	Mitsubishi Electric	University of Tokyo
19	Basel	CH/DE/FR	DSM IP Assets	University of Basel
20	Copenhagen	DK	Novo Nordisk	University of Copenhagen
21	Nuremberg–Erlangen	DE	Siemens	University of Erlangen Nuremberg
22	Stuttgart	DE	Robert Bosch	Eberhard Karls University of Tübingen
23	Minneapolis, MN	US	3M Innovative Properties	University of Minnesota Twin Cities
24	Pittsburgh, PA	US	University of Pittsburgh	University of Pittsburgh
25	Seoul	KR	Samsung Electronics	Seoul National University

Quelle: https://www.wipo.int/tech_trends/en/artificial_intelligence/story.html

Companies represent 26 of the world's top 30 AI patent applicants

1

Private GPT protects the company's IP

2

Pay-per-use OnPrem instead of cost risk

3

High performance and highly scalable



Q & A

