

10 June 2026

The XR5.0 LLM Engine: RAG at the Edge of XR Industrial Work

George Fatouros | INNOV-ACTS

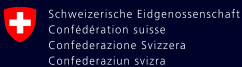
XR5.0

Human-Centric AI-Enabled
Extended Reality Applications
for the Industry 5.0 Era



Funded by
the European Union

Project funded by



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI

Swiss Confederation

This work has received funding from the Swiss
State Secretariat for Education, Research and
Innovation (SERI)

Alarm 504. Live machine. Two hands occupied.

"A bar-feeder throws Alarm 504 at a steel manufacturer's facility. The technician is working hands-free on the live machine. The corrective procedure is somewhere in a 300-page manual — in a cabinet across the room. Every minute of downtime costs the line."

TACIT EXPERTISE

- Scarce specialists, often unavailable remotely
- Experience is hard to scale across shifts and sites

TECHNICAL DOCUMENTATION

- Long manuals, visual procedures, poor searchability
- Inaccessible at the point of work, under time pressure

Industrial knowledge-access failure translates directly into downtime, errors, and safety risk.



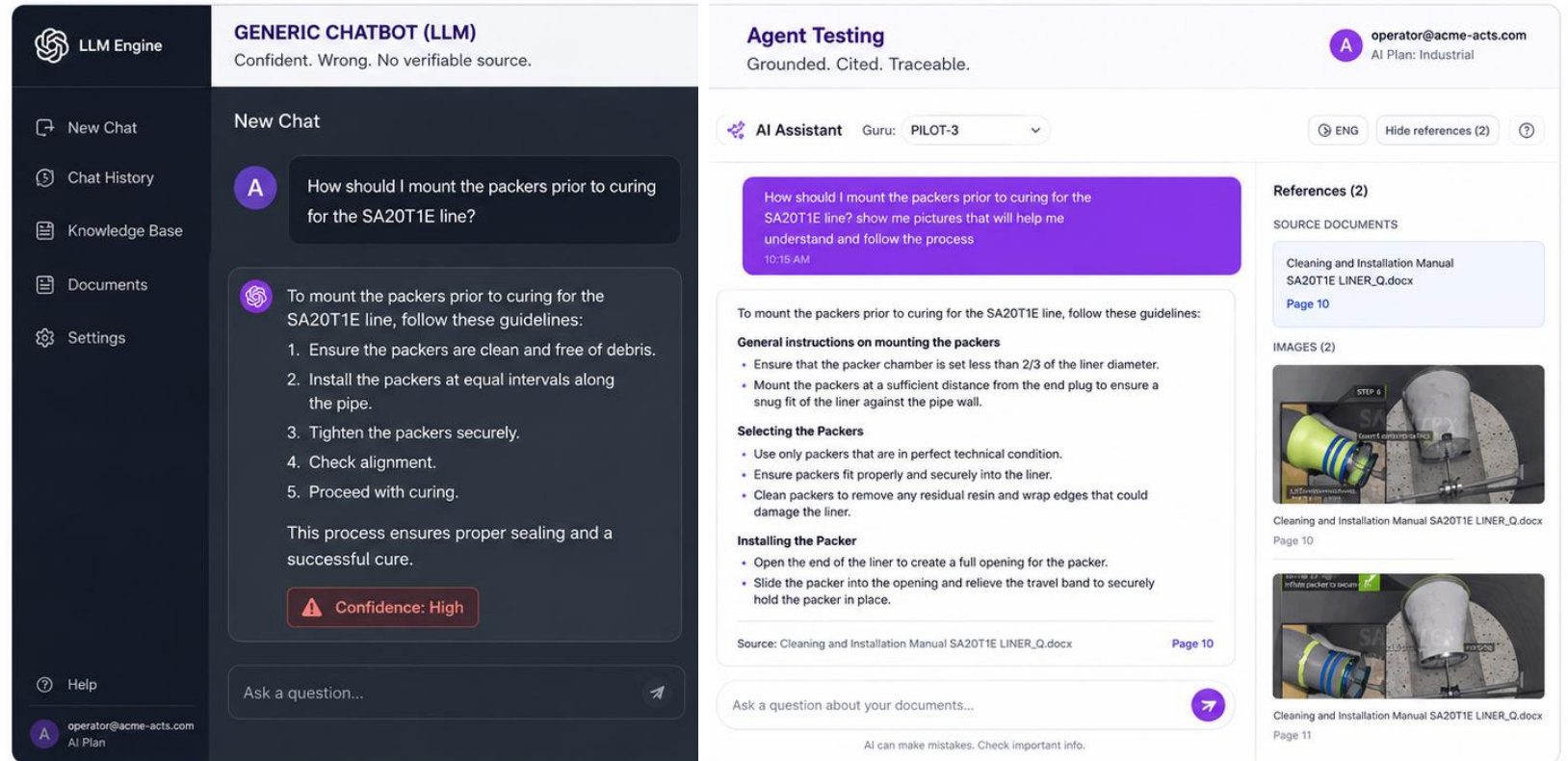
A foundational AI model is not an industrial assistant

A FLUENT LLM

- General pre-training knowledge
- No access to your documentation
- No verified source references
- No XR integration

AN INDUSTRIAL ASSISTANT

- Answers grounded in approved documents
- Scoped retrieval with page-level citations
- Abstains rather than fabricates
- Adapt to the end user



GENERIC CHATBOT (LLM)
Confident. Wrong. No verifiable source.

Agent Testing
Grounded. Cited. Traceable.

Generic Chatbot Response:

How should I mount the packers prior to curing for the SA20T1E line?

To mount the packers prior to curing for the SA20T1E line, follow these guidelines:

1. Ensure the packers are clean and free of debris.
2. Install the packers at equal intervals along the pipe.
3. Tighten the packers securely.
4. Check alignment.
5. Proceed with curing.

This process ensures proper sealing and a successful cure.

Confidence: High

Agent Testing Response:

How should I mount the packers prior to curing for the SA20T1E line? show me pictures that will help me understand and follow the process

To mount the packers prior to curing for the SA20T1E line, follow these guidelines:

General instructions on mounting the packers

- Ensure that the packer chamber is set less than 2/3 of the liner diameter.
- Mount the packers at a sufficient distance from the end plug to ensure a snug fit of the liner against the pipe wall.

Selecting the Packers

- Use only packers that are in perfect technical condition.
- Ensure packers fit properly and securely into the liner.
- Clean packers to remove any residual resin and wrap edges that could damage the liner.

Installing the Packer

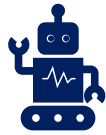
- Open the end of the liner to create a full opening for the packer.
- Slide the packer into the opening and relieve the travel band to securely hold the packer in place.

Source: Cleaning and Installation Manual SA20T1E LINER_Q.docx Page 10

References (2)
SOURCE DOCUMENTS
Cleaning and Installation Manual SA20T1E LINER_Q.docx Page 10
IMAGES (2)
Cleaning and Installation Manual SA20T1E LINER_Q.docx Page 10
Cleaning and Installation Manual SA20T1E LINER_Q.docx Page 11

Real failure — a pilot LLM correlated a page header numbered “8” with a sensor count (“8 LASER DISTANCE SENSORS”) and produced a confident, wrong answer. The operator had no way to detect the error without checking the manual.

A reusable knowledge layer for human-centred industrial XR



AGENTIC

Plans what to retrieve, how many searches to issue, and which documents to target — before generating a single word of response.



MULTIMODAL

Preserves figures, schematics, and tables at ingestion. Technical images are described, filtered, and linked to the text chunks they illustrate — retrievable alongside text.



USE-CASE AGNOSTIC

One deployment serves unrelated industrial domains through configuration alone. Pilot knowledge is isolated; the pipeline is shared.

How Documents Become Queryable: The Ingestion Pipeline

Industrial documents are multimodal — the ingestion pipeline preserves that

Most RAG pipelines extract text and discard everything else.

This pipeline treats figures and tables as first-class retrievable content.

Each image is described by a vision-language model, filtered for technical relevance (logos and decorative graphics are dropped), and **bound to the text chunk it illustrates at ingestion time** — before any query is ever issued.

Each use case is stored in its own isolated vector namespace.

1 Structure-preserving OCR

Headings · sections · tables · figures



2 Image Filtering & Description

Drops logos / decorative graphics; VLM generates technical descriptions



3 Segment-aware Chunking

Chunks carry section metadata + figure references



4 Embedding & Storage

Vector embeddings → Vector DB namespace per use case

Image references travel with their chunk. When a chunk is retrieved, its visual context is already attached.

From a Question to a Grounded Answer: The Query Pipeline

Three stages, one short-lived workflow per request

1 — RETRIEVAL PLANNING

The tool-planning LLM receives the query, session history, use-case system prompt, and a catalogue of available documents with summaries. It decides whether retrieval is needed, and if so, issues one or more targeted vector searches — letting multi-part questions be handled by several focused searches rather than one broad one.

2 — IMAGE SELECTION

Retrieved chunks arrive with their associated figure references already attached. A second LLM selects only the figures needed to understand the answer. Marginal images never reach the response.

3 — ANSWER GENERATION

A third LLM receives the use-case system prompt, the retrieved text, and a summary of the selected figures. It produces a grounded response with page-level citations and explicitly states where the documentation does not cover the question.

A short-lived workflow runs for every request. No state persists after response. Use-case isolation is enforced by data partitioning, not separate deployments.

Tool-Planning LLM

↓ query + docs catalogue → tool calls

Image Selection LLM

↓ retrieved chunks + figure refs → filtered figures

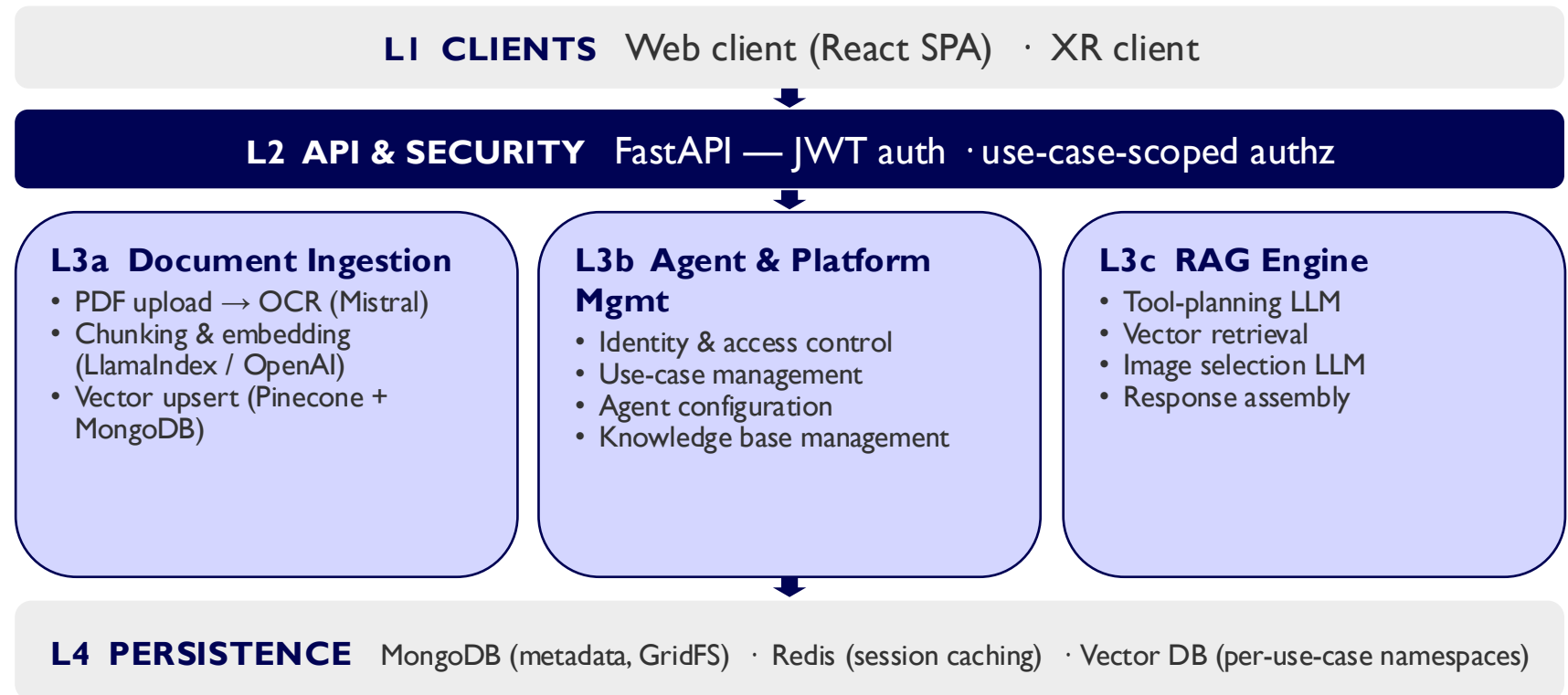
Answer Generation LLM

↓ grounded response + page references

Four layers connect clients to secured persistence

Every interaction crosses the same authentication and use-case boundary. No client or model reaches storage directly.

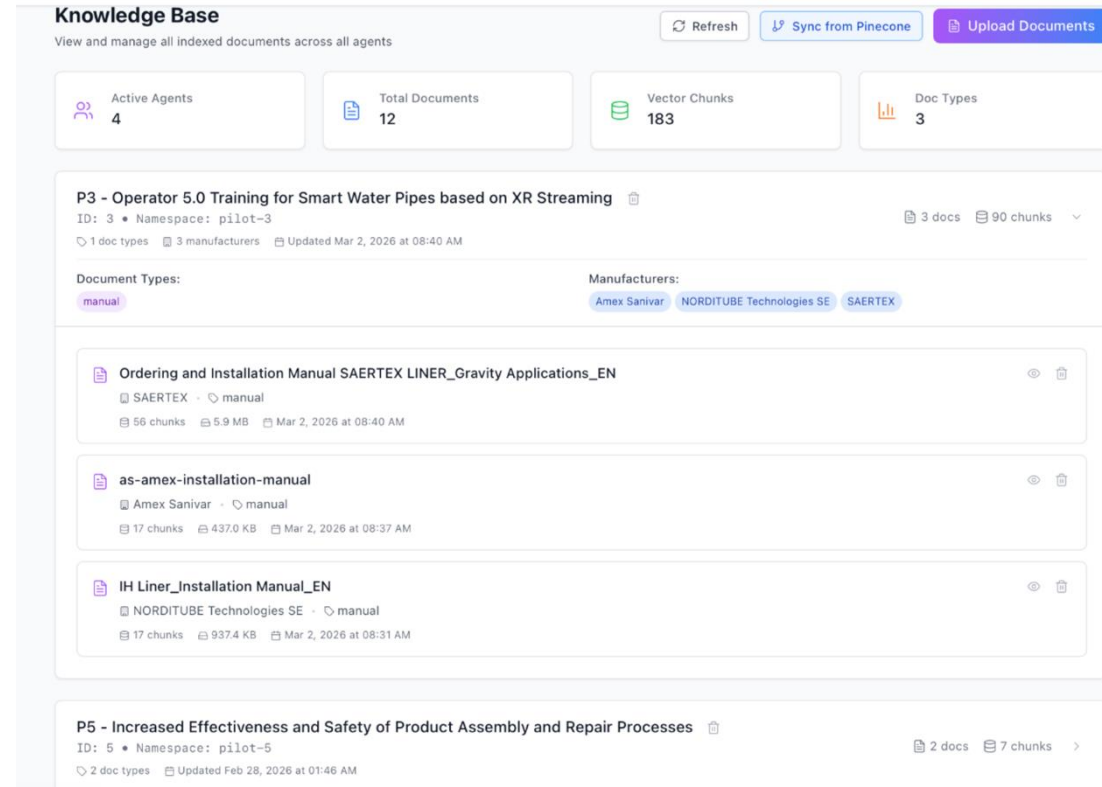
The system is **model-agnostic**: OCR, embedding, VLM captioning, and inference LLMs are each consumed behind an abstraction and selected through configuration — commercial API in one setting, on-premise stack in another.



Stack: FastAPI · Mistral OCR · Mistral Pixtral (VLM) · OpenAI text-embedding-3-small · GPT-4.1-mini · Pinecone · MongoDB/GridFS · Redis

One Platform for Different Industrial Use Cases

Pilot	Context	What the agent does
Pilot 3 (EKSO)	Smart water pipe inspection	Installation procedure guidance with visual aids
Pilot 5 (SPACE)	Edge device assembly training	Self-service VR/AR training support
Pilot 6 (LNS)	Industrial machine troubleshooting	Alarm diagnosis + corrective action retrieval



Knowledge Base
View and manage all indexed documents across all agents

Refresh Sync from Pinecone Upload Documents

Active Agents: 4 | Total Documents: 12 | Vector Chunks: 183 | Doc Types: 3

P3 - Operator 5.0 Training for Smart Water Pipes based on XR Streaming
ID: 3 • Namespace: pilot-3 | 3 docs | 90 chunks | Updated Mar 2, 2026 at 08:40 AM

Document Types: manual | Manufacturers: Amex Sanivar, NORDITUBE Technologies SE, SAERTEX

- Ordering and Installation Manual SAERTEX LINER_Gravity Applications_EN (56 chunks, 5.9 MB, Mar 2, 2026 at 08:40 AM)
- as-amex-installation-manual (17 chunks, 437.0 KB, Mar 2, 2026 at 08:37 AM)
- IH Liner_Installation Manual_EN (17 chunks, 937.4 KB, Mar 2, 2026 at 08:31 AM)

P5 - Increased Effectiveness and Safety of Product Assembly and Repair Processes
ID: 5 • Namespace: pilot-5 | 2 docs | 7 chunks | Updated Feb 28, 2026 at 01:46 AM

Pilot-agnostic architecture: A single deployment, independently configured per use case via system prompt + knowledge base. Self-service Integration

Evaluation: Measured at the Claim Level

RAGChecker across 40 operational queries, multiple use cases

Evaluation uses **RAGChecker**, which decomposes each response into individual claims and compares them against ground-truth answers — separating retrieval failures from generation failures.

40 queries span heterogeneous document types: comprehensive installation/inspection manuals (Pilot 3) and concise alarm-code procedures (Pilot 6).

Metric	Mean
Faithfulness	97.7%
Hallucination	2.0%
Context Utilisation	75.2%
Claim Recall	84.4%
Context Precision	66.0%

Metric	Pilot 3 (EKSO) k=5	Pilot 6 (LNS) k=7
Faithfulness	0.929	0.980
Hallucination	0.037	0.011
Context Utilisation	0.763	0.700
Claim Recall	0.854	0.950
Context Precision	0.570	0.700

Cross-pilot comparison confirms reusability with numbers: different document structures, different sectors — same pipeline, similar or better performance in Pilot 6.

Trustworthy behaviour is an architectural property, not a prompt trick

INTEGRATION

- REST API — web and XR clients consume the same endpoints
- Per-request use-case scoping — no cross-pilot knowledge bleed
- Stateless workflows — horizontally scalable, no session stickiness

TRUST

- JWT authentication + use-case-scoped authorisation at a single API boundary
- Source references with page numbers in every response
- **Abstention over hallucination:** outside the ingested documentation, the system says so rather than generating an unsupported answer
- Multi-turn coherence: follow-ups interpreted in context of prior text and images

Chat		
POST	/api/chat	Chat
DELETE	/api/chat/history	Clear Chat History
Documents		
GET	/api/documents/overview	Get Documents Overview
GET	/api/documents/pilot/{pilot_id}	Get Pilot Documents
DELETE	/api/documents/pilot/{pilot_id}	Delete Pilot Namespace
GET	/api/documents/my-documents	Get My Documents
GET	/api/documents/file/{collection_name}	Get Source Document File
DELETE	/api/documents/pilot/{pilot_id}/document/{collection_name}	Delete Document
DELETE	/api/documents/cleanup-empty-namespaces	Cleanup Empty Namespaces
POST	/api/documents/sync-from-pinecone	Sync From Pinecone
GET	/api/documents/dashboard/documents	Get Dashboard Documents
Document Upload		
POST	/api/upload	Upload Document
System Prompts		
GET	/api/system_prompt/all	List System Prompts
POST	/api/system_prompt/get	Get System Prompt
POST	/api/system_prompt/	Upsert System Prompt
DELETE	/api/system_prompt/	Delete System Prompt

Latency is the known cost of the three-stage pipeline. It is monitored and will be addressed in future work — not hidden.

Live Production Demo

XR5.0

Human-Centric AI-Enabled
Extended Reality Applications
for the Industry 5.0 Era

Contact: gfatouros@innov-acts.com

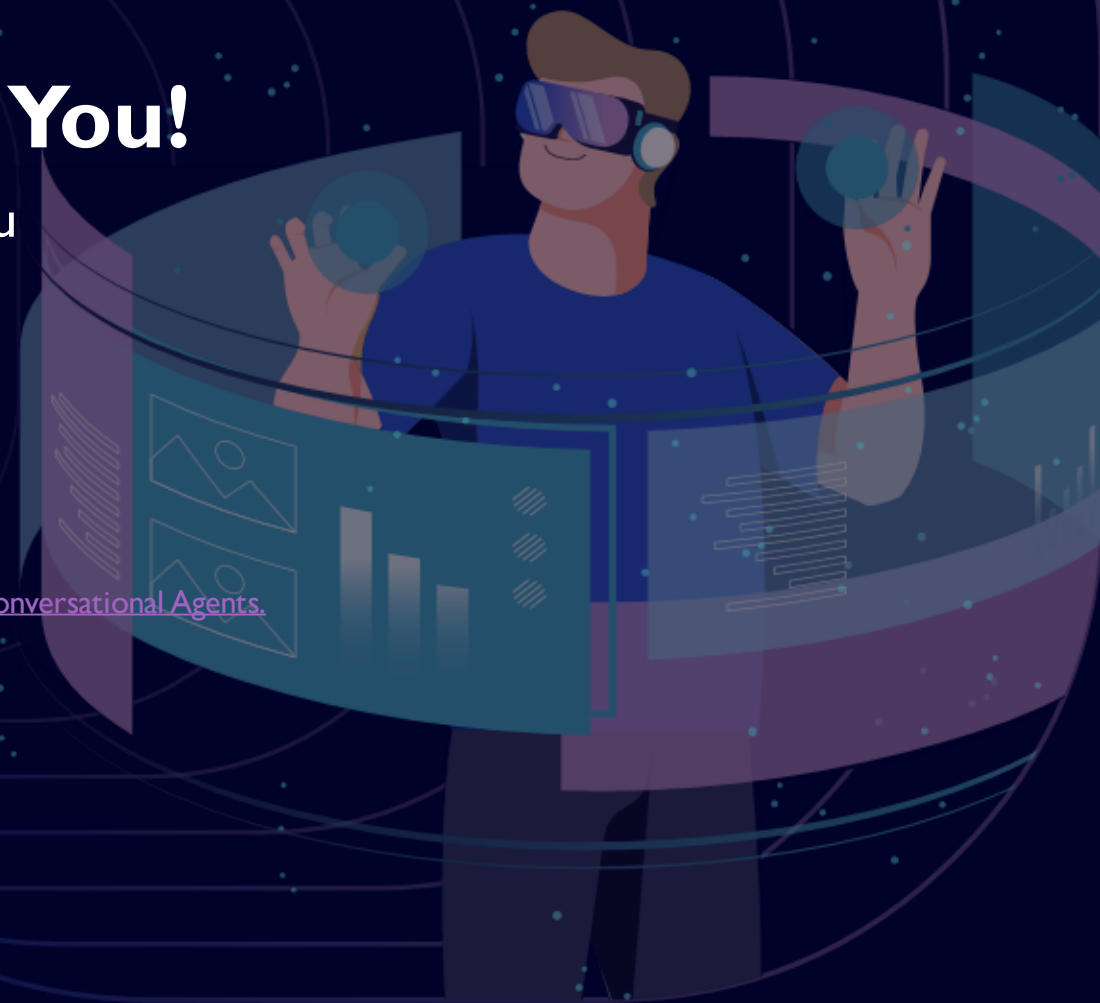
Platform: <https://chat.xr50.eu>

API Docs: chat-api.xr50.eu/docs

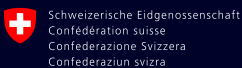
Paper: [Tomkou, D., Fatouros et al. \(2025\). Bridging Industrial Expertise and XR with LLM-Powered Conversational Agents.](#)

Thank You!

www.xr50.eu



Project funded by



Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI

This work has received funding from the Swiss
State Secretariat for Education, Research and
Innovation (SERI)