



How GenAI improves information extraction from complex industrial documents

29 June 2026, 12:00

Sreeraj Rajendran

Nicolás González-Deleito

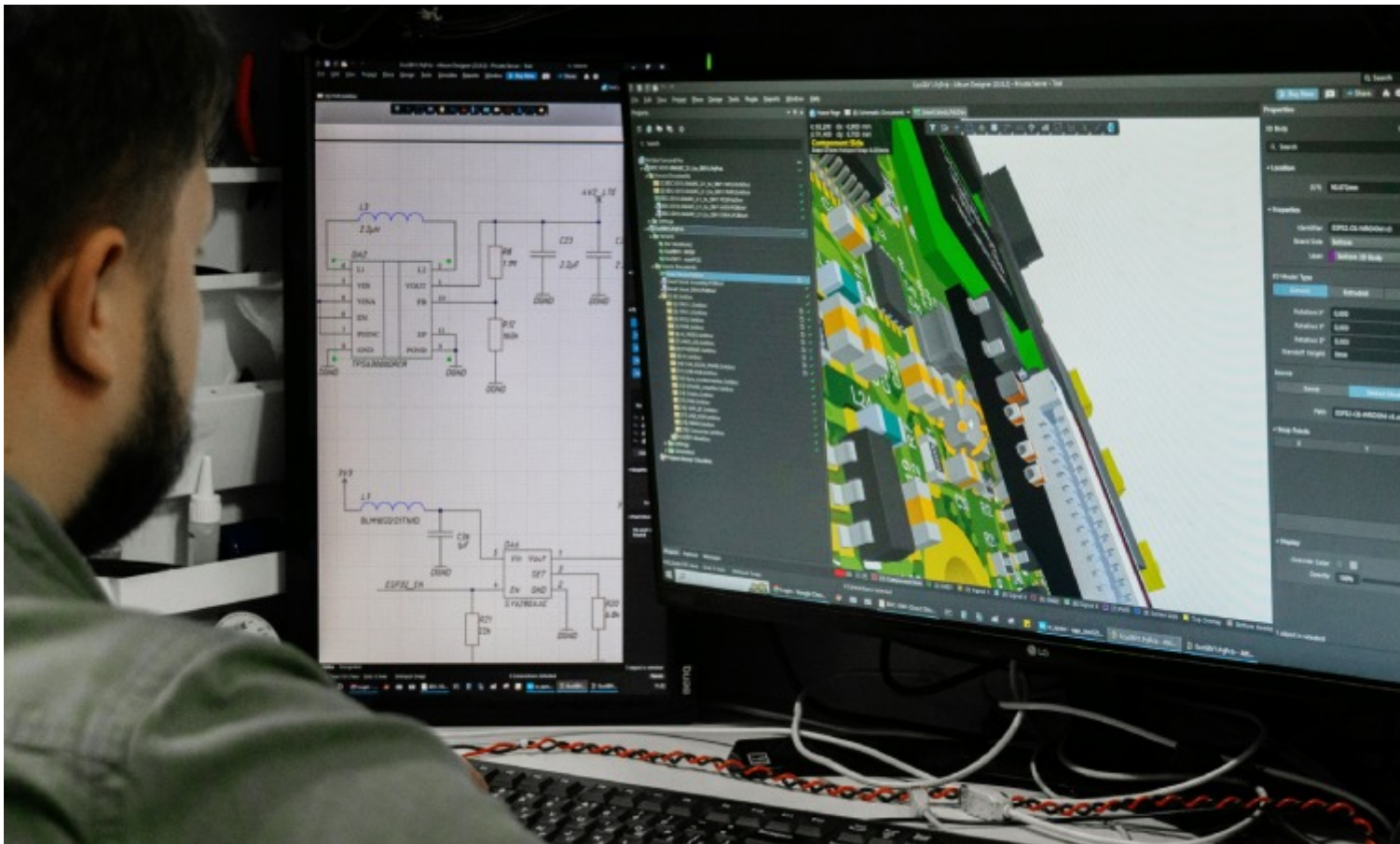
Ali Beigrezaei

Aziz Nebli

Turning technical documents into actionable information

Industrial processes, factory layouts, plant configurations, and component specifications are typically documented in a wide range of technical sources: textual descriptions, schematics, figures, tables, and images. Extracting relevant information from these documents is essential for activities such as maintenance, audits, certifications, and digital twin creation. However, reliably identifying, linking, and interpreting information across these heterogeneous sources remains a complex task.

Through the international [NARRATE](#) project, supported by the Brussels Capital Region and Innoviris, Sirris and 3E are exploring how Generative AI can automatically extract, connect, and structure this information, paving the way for faster, more efficient processing and automation workflows.



About NARRATE

[NARRATE](#) is an international research project bringing together 14 partners from five countries: Belgium, Estonia, South Korea, Spain, and Türkiye. The project focuses on developing trustworthy and ethical AI-driven human interaction interfaces, combining multi-modal data integration with automated knowledge updates.

[Read more](#)

Why information extraction remains difficult

The complexity does not lie in finding individual pieces of information, but in understanding how they are all connected across different sources:

- Input is often unstructured and distributed via different formats, such as text, tables, and images

- Information is linked between formats, such as a text description referring to information in a table or image
- References are located far apart within a single document or across multiple files
- The same concept can be described using different naming conventions or identifiers

As a result, extracting information is only part of the problem. Systems must also identify relationships, consolidate information, and maintain context across heterogeneous document collections.

Turning heterogeneous documents into usable data

Within [NARRATE](#), Sirris and 3E are developing AI-driven pipelines that can automatically extract, connect, and structure information from a wide range of engineering documents. To achieve this, Sirris and 3E combine two complementary AI strategies.

1. Local LLMs

Sirris focuses on locally hosted language models (LLMs) such as Llama, Gemma, and Mistral. Keeping models on internal infrastructure ensures that sensitive customer data remains within the company environment. The initial workflow follows a multi-step approach.

1. Identify document type and content
2. Extract relevant information
3. Structure the extracted data
4. Repeat the above steps for each document
5. Consolidate the information from different files into a single output

As the project progressed, linking information across multiple documents proved particularly challenging. This led to a shift towards an agent-based workflow, where AI coordinates the different steps required to detect, connect, and validate information across an entire document set.

This agentic workflow enables the system to handle versatile, multimodal document sets containing a mix of text, tables, figures, and schematics across varying formats and naming conventions.

2. Cloud-based LLMs

In parallel, 3E is developing a cloud-based pipeline within a secure Azure environment. Documents are first processed and intelligently segmented using technologies such as Azure Computer Vision and Document Intelligence. GenAI models then identify information types and attempt to automatically reconstruct relationships between them. The objective is the same: to transform complex engineering documentation into a structured consolidated view, with as little manual intervention as possible.



Improving reliability through validation

To improve reliability, both partners developed extensive validation mechanisms. Sirris uses Langfuse to monitor prompts, compare model parameters, and evaluate output quality. 3E applies an ensemble approach in which multiple LLMs perform the same extraction task in parallel. The resulting outputs are then cross-checked for inconsistencies, missing information, and duplicate information before final consolidation.

How do you measure success?

One important question remained: how do you evaluate the quality of automatically extracted information? The two partners developed custom evaluation metrics focusing on:

- Completeness
- Presence of correct items
- Field quality
- Numerical correctness

The first results are promising. Information extraction already performs strongly. Automatically linking complex relationships between information items, however, remains the most difficult part.

What are the next steps?

The first phase of the project has already demonstrated that GenAI can meaningfully accelerate the extraction of technical information from complex engineering documentation. The next step is to extend these capabilities further by enabling complex component linking and translating the extracted knowledge into practical decision support. The consortium will also explore applications such as intelligent dashboards, recommendation systems, and automated knowledge management for operational decision support.

The journey is far from over. However, the progress achieved within [NARRATE](#) already demonstrates how complex engineering documentation can be transformed into actionable digital intelligence. As GenAI continues to mature, reliable information extraction could become a key enabler for improving processes, such as faster onboarding, greater scalability, and more efficient asset management.

Curious how GenAI could help streamline your own engineering workflows?

Sirris helps companies apply AI and GenAI to complex industrial challenges, from information extraction and knowledge management to automation and decision support.

[Contact us](#)

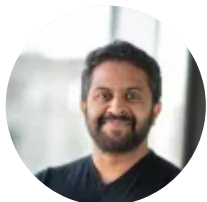
Partners

with the support of ITEA4, Innoviris and the Brussels-Capital Region



Co-funded by

Authors



Sreeraj Rajendran



Nicolás González-Deleito



Ali Beigrezaei



Aziz Nebli