Hutnické listy č. 1/2026, roč. LXXIX
ISSN Online: 3029-8350

Artificial Intelligence Tools in Industrial Control: Large Language Models, Agentic Systems and Their Current Deployment

Nástroje umělé inteligence v průmyslovém řízení: Velké jazykové modely, agentní systémy a jejich současné využití

Ing. Dagmar Špičková, Ph.D.[1]; doc. Ing. Ivo Špička, Ph.D.[2]

[1] VŠB-TUO, FMT, 17. listopadu 2172/15, 708 00 Ostrava – Poruba, Czech Republic, dagmar.spickova@vsb.cz
[2] BINTELL SOLUTIONS, Na Vyhlídce 465, 742 85 Vřesina, Czech Republic, ivo.spicka@gmail.com

Abstract

This paper follows on from last year's review of artificial intelligence tools in industrial control and updates it with key trends from 2025. A fundamental shift has occurred in the area of large language models (LLMs), which have become a standard part of the enterprise environment – notably models GPT-4o, Claude 3.7, Gemini 2.0 and DeepSeek R1. Concurrently, there has been a massive expansion of agentic AI systems, which – unlike conversational chatbots – autonomously plan, reason and execute multi-level workflows. The paper describes key differences between cloud-based and on-premise LLM deployment, approaches to building custom enterprise language systems (fine-tuning, RAG, private models), and methods for automated processing of corporate technical documentation. Specific industrial deployments at companies such as Siemens, ABB and SMS Group are discussed. The paper concludes with perspectives for heavy industry and steelmaking in the context of the EU AI Act regulatory framework.

Key words: *large language models, agentic AI, RAG, on-premise LLM, technical documentation*

Abstrakt

Tento článek navazuje na loňský přehled nástrojů umělé inteligence v průmyslovém řízení a aktualizuje jej o klíčové trendy roku 2025. Zásadní posun nastal v oblasti velkých jazykových modelů (LLM), které se staly standardní součástí podnikového prostředí. Mezi nejvýznamnější patří zejména modely GPT-4o, Claude 3.7, Gemini 2.0 a DeepSeek R1. Současně došlo k výraznému rozvoji agentních systémů umělé inteligence, které na rozdíl od konverzačních chatbotů dokážou autonomně plánovat, uvažovat a vykonávat víceúrovňové pracovní postupy. Článek popisuje hlavní rozdíly mezi nasazením LLM v cloudovém prostředí a v režimu on-premise, přístupy k budování podnikových jazykových systémů na míru (fine-tuning, RAG, privátní modely) a metody automatizovaného zpracování podnikové technické dokumentace. Dále jsou diskutovány konkrétní průmyslové implementace ve společnostech Siemens, ABB a SMS Group. V závěru jsou nastíněny perspektivy využití těchto technologií v těžkém průmyslu a hutnictví v kontextu regulačního rámce EU AI Act.

Klíčová slova: *velké jazykové modely, agentní umělá inteligence, lokální LLM, technická dokumentace*

1. Introduction

The previous paper [1] comprehensively mapped the main AI tools in industrial control – machine learning, computer vision, digital twins, expert systems and natural language processing (NLP) – and demonstrated their application potential in the automotive, process and energy industries. Since then, the technological landscape has changed fundamentally.

Hutnické listy č. 1/2026, roč. LXXIX
ISSN Online: 3029-8350

The global industrial AI market reached a value of USD 43.6 billion in 2024; analysts forecast annual growth of approximately 23% to USD 153.9 billion by 2030 [2]. A breakthrough has occurred in two areas in particular: (1) large language models (LLMs) have become an accessible tool for enterprises of all sizes; (2) agentic AI systems have moved from passively answering queries to actively and autonomously executing complex industrial tasks [3].

This paper aims to update the previous review with these new AI dimensions, with emphasis on three practical areas: the choice of deployment model (cloud vs. on-premise), the creation of custom enterprise LLM systems, and automated processing of technical documentation.

2. Large Language Models – State of the Art in 2025

2.1 Key models and their industrial relevance

A large language model (LLM) is a neural network based on the Transformer architecture, trained on billions of text tokens, capable of generating, translating, summarising and analysing natural language. The year 2025 brought several generational transitions [4]:

- **GPT-4o and the "o" model series (OpenAI):** Natively multimodal models processing text, images and audio. The "o" series introduced advanced chain-of-thought reasoning, suitable for complex technical tasks.
- **Claude 3.7 Sonnet (Anthropic):** A model excelling in complex reasoning and document analysis; preferred for industrial applications requiring high reliability.
- **Gemini 2.0 / 2.5 Pro (Google DeepMind):** Natively multimodal models with a context window of up to 1 million tokens, well suited for working with extensive technical documentation.
- **DeepSeek R1 (DeepSeek AI):** An efficient open-source model with an excellent performance-to-cost ratio, deployable on-premise without dependence on cloud providers.

A 2025 survey found that 35% of respondents from industrial enterprises had already actively deployed LLMs or AI agents in their operational environment [5].

3. Cloud vs. On-Premise LLM Deployment in Industry

3.1 Cloud deployment

Cloud-based LLM deployment (via API services such as OpenAI, Google Vertex AI or Azure OpenAI) currently represents the dominant model, accounting for approximately 65% of the enterprise LLM market [6]. The main advantages lie in immediate availability without hardware investment, automatic model updates, easy scalability and access to the most capable available models.

The critical disadvantage for industrial enterprises is the transfer of sensitive operational data (production parameters, recipes, metallurgical composition) outside the corporate network to a third-party provider. Further concerns include dependence on internet connectivity, latency unsuitable for time-critical tasks, and ongoing per-token costs that can be substantial under intensive industrial use [7].

Hutnické listy č. 1/2026, roč. LXXIX
ISSN Online: 3029-8350

3.2 On-premise deployment

On-premise (local) LLM deployment means running the model directly on the enterprise's own servers or on industrial edge hardware. The key enabler of this approach is the emergence of compact Small Language Models (SLMs) – Microsoft Phi-3, LLaMA 3, Mistral – with parameter counts in the range of a few billion, which achieve performance comparable to previous-generation large models and can be operated on industrial hardware without top-tier GPU accelerators [8].

Advantages of on-premise deployment:

- **Data protection:** Sensitive operational and business data never leaves the corporate network.
- **Low latency:** Response times in the millisecond range, suitable for real-time industrial control.
- **Regulatory compliance:** Fulfilment of data localisation requirements (GDPR, sector-specific regulations).
- **Predictable costs:** One-time hardware investment with no ongoing API usage fees.

Disadvantages include higher initial investment, the need for internal management, and generally lower performance compared to the latest cloud models [9].

3.3 Hybrid architecture as the industrial standard

The most prevalent approach in industry in 2025 has become a hybrid architecture in which cloud and on-premise deployments complement each other (**tab. 1**) [10]:

Tab. 1 Comparison of cloud, on-premise and hybrid LLM deployment in industry
Tab. 1 Srovnání nasazení velkých jazykových modelů (LLM) v cloudu, na místě a v hybridním režimu v průmyslu

Parameter	*Cloud LLM*	*On-Premise LLM*	*Hybrid*
Data security	*Lower (data off-site)*	*High*	*High (sensitive data local)*
Latency	*Higher (ms–s)*	*Low (< 10 ms)*	*Optimised*
Model performance	*Highest*	*Medium*	*High*
Initial cost	*Low*	*High*	*Medium*
Operational cost	*Ongoing (per token)*	*Low*	*Combined*
Scalability	*High*	*Limited*	*High*
Suitability for real-time	*Limited*	*High*	*High*

In a hybrid architecture, the local SLM handles latency-sensitive tasks (anomaly detection, real-time process control), while the cloud model performs computationally intensive analyses (model training, complex document analysis). Siemens has implemented this strategy in its Industrial Copilot platform, where edge models communicate with the Xcelerator cloud environment [11].

4. Building Custom Enterprise LLM Systems

4.1 The spectrum of approaches: from prompting to training a custom model

Enterprises have access to a full spectrum of approaches for building their own AI systems, differing in cost, complexity and degree of customisation [12]:

Hutnické listy č. 1/2026, roč. LXXIX
ISSN Online: 3029-8350

(1) Prompt engineering and in-context learning: The fastest and least expensive approach – an existing model is controlled through carefully designed instructions and examples in the prompt. Suitable for standardised, repetitive tasks such as report generation or incident classification.

(2) Retrieval-Augmented Generation (RAG): The model is connected to the enterprise data repository via a vector database. Each query is first used to retrieve relevant documents, which are then passed to the model as context. RAG significantly reduces hallucinations and enables the model to work with up-to-date enterprise data without retraining [13]. RAG is currently the most widely adopted approach for industrial applications.

(3) Fine-tuning: An existing pre-trained model is further trained on an enterprise dataset (technical standards, operational protocols, fault records, metallurgical databases). The result is a model specialised in the enterprise's domain with improved accuracy for domain-specific topics. The key technique is LoRA (Low-Rank Adaptation), which enables efficient fine-tuning even on relatively small datasets [14]. A 2025 analysis reports that 70–80% of enterprise LLM use cases are better served by fine-tuning existing models than by the costly process of training from scratch [15].

(4) Training a custom model from scratch: The costliest and technically demanding approach, reserved for enterprises with uniquely specific needs or exceptionally large proprietary datasets. Siemens is investing in this approach with its Industrial Foundation Model project – a domain-specific generative AI model trained on industrial data across its entire ecosystem [11].

4.2 Data preparation as a critical factor

The quality of the resulting enterprise LLM system is directly dependent on the quality of the input data. Industrial datasets are typically characterised by high heterogeneity: structured data from SCADA/MES systems, unstructured textual documentation, CAD drawings and audio records from maintenance inspections. Key data preparation steps include:

Cleaning and deduplication of historical documents (particularly in archives of older standards and drawings),

Annotation of enterprise terminology and creation of a domain vocabulary (ontology),

Versioning of documentation to ensure context currency,

Anonymisation of sensitive business information before use in cloud services.

5. Automated Processing of Enterprise Technical Documentation

5.1 Challenges of industrial documentation

Industrial enterprises, and steelworks in particular, manage an extensive ecosystem of technical documentation: operational procedures, standards (EN, ISO, ASTM), drawings, incident reports, heat records, material certificates, inspection reports, and maintenance logs.

This documentation is typically stored in fragmented systems, in various formats (PDF, Word, paper scans) and in multiple languages. Surveys indicate that industrial technicians spend an average of 20–30% of their working time searching for and verifying information in documentation [16].

Hutnické listy č. 1/2026, roč. LXXIX
ISSN Online: 3029-8350

5.2 RAG system architecture for technical documentation

The fundamental building block of a modern enterprise documentation system is the RAG architecture complemented by a vector database [13]. The process involves two parallel branches (**fig. 1**):

Document indexing encompasses the ingestion and OCR conversion of documents in various formats, splitting them into semantic blocks (chunking), vectorising the blocks using an embedding model, and storing the vectors in a specialised database (Chroma, Pinecone, Milvus).

Querying and answer generation works as follows: the technician's natural-language query is also vectorised, the system performs a semantic search for the most relevant blocks in the database, these blocks are passed to the LLM as context, and the model generates an answer with a reference to the specific source document and page number.

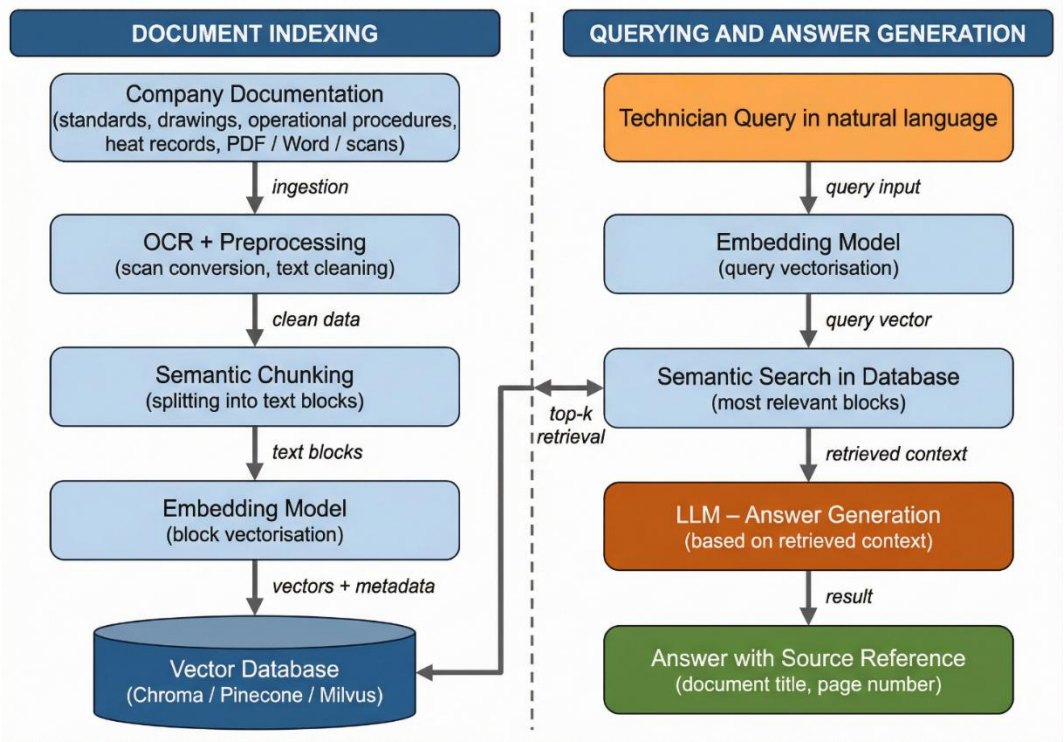

Fig. 1 RAG system architecture for processing industrial technical documentation
Obr. 1 Architektura systému RAG pro zpracování průmyslové technické dokumentace

Implementations of RAG have been shown to reduce the frequency of incorrect or incomplete answers by 40–50% compared with pure LLMs [13].

5.3 Specific applications in industrial practice

Intelligent search in standards and regulations: A technician submits a query in natural language and the system immediately returns a precise answer with a reference to the specific standard and page.

Historical fault analysis: The RAG system searches the fault protocol database and identifies whether a similar fault has previously occurred, what the root cause was and how it was resolved. Knowledge management systems built on operational logs have been shown to reduce the resolution time for recurring problems by 60–75% [17].

Hutnické listy č. 1/2026, roč. LXXIX
ISSN Online: 3029-8350

Operational documentation generation: Agentic systems automatically generate heat protocols, quality inspection reports or maintenance records based on structured data from SCADA/MES systems, significantly reducing the administrative burden on operators.

Multilingual documentation: For multinational steel groups, LLMs automate the translation of technical documentation while preserving domain-specific terminology – a task at which general-purpose translation tools fail.

Onboarding of new employees: The RAG system acts as an interactive guide for new workers, who can ask questions about technological procedures and safety regulations in natural language, without having to read hundreds of pages of documentation.

5.4 Case study: SMS Group

SMS Group has implemented an intelligent technical documentation management and process report automation system based on LLMs in its customer steelworks. The system encompasses autonomous recommendation of heat parameters, quality prediction and real-time operational documentation generation [18]. According to published results, the system has reduced the time needed to access relevant technical information by more than 60% and has decreased the number of errors caused by incorrect or outdated documentation.

6. Agentic AI Systems – A Breakthrough in Industrial Control

6.1 Principles of agentic AI

Agentic AI represents a qualitative leap beyond conversational LLMs. While a conventional chatbot responds to individual queries, an AI agent autonomously: (1) plans a sequence of steps to achieve a given goal; (2) decides which tools to use (databases, APIs, computational modules, sensor data); (3) executes actions in a real or simulated environment; (4) iterates based on feedback and new information [3].

The key concept is Multi-Agent Systems (MAS), in which multiple specialised agents collaborate: one monitors sensor data, another searches technical documentation, a third communicates with the ERP system, and a fourth drafts and dispatches work orders. The agentic AI market in manufacturing reached USD 5.5 billion in 2025, growing at approximately 25% per year [19].

6.2 Industrial deployments

Siemens presented its Industrial AI Agents platform at the Automate 2025 trade show, integrating agents into the SIMATIC and Xcelerator ecosystem. Agents autonomously diagnose production line faults, generate service orders and plan maintenance without direct operator intervention [11].

ABB, in partnership with Microsoft, launched the Genix Copilot application built on Azure OpenAI, combining LLMs with the ABB Ability™ data platform. The system analyses the performance of energy-intensive industrial equipment and recommends optimisation interventions [20].

SMS Group has deployed AI agents for automated process control in steelworks, encompassing autonomous heat parameter recommendation, quality prediction and real-time operational documentation generation [18].

Hutnické listy č. 1/2026, roč. LXXIX
ISSN Online: 3029-8350

6.3 Perspectives for agentic AI in steelmaking (tab. 2)

Tab. 2 Applications of agentic AI in the steelmaking industry
Tab. 2 Využití agentní umělé inteligence v ocelářském průmyslu

Area	***Description of agentic solution***	***Expected benefit***
EAF heat control	*Agent monitors spectrometric data, scrap history and energy tariff; recommends charge composition and process parameters in real time*	*Reduction of electricity consumption by 3–8%*
Predictive maintenance	*Multi-agent system collects vibration, temperature and fault history data; autonomously orders spare parts and schedules maintenance windows*	*Reduction of unplanned downtime by 20–35%*
Quality control	*Agent processes data from measuring lines and compares with EN and ASTM standards; initiates heat regrading or scrapping*	*Reduction of customer complaints*
Documentation management	*RAG agent answers technician queries from current drawings, standards and operational procedures*	*Reduction of information retrieval time by 60%*
Logistics optimisation	*Agent coordinates scrap orders, production planning and finished product deliveries*	*Inventory and transport optimisation*

7. Industry 5.0, EU AI Act and Future Perspectives

Agentic AI systems are the natural technological expression of the Industry 5.0 paradigm: agents take over routine cognitive tasks while the human expert focuses on creative decision-making and exception handling. In steelmaking, this means that an EAF operator becomes a "supervisor of AI agents" rather than a manual controller of process parameters. LLMs function best as an extension of human expertise, not as its replacement [21].

The entry into force of the EU AI Act (August 2025) has introduced new obligations for operators of AI systems in industry. Industrial AI systems affecting worker safety are classified as "high-risk" and are subject to requirements for transparency, auditability and human oversight. This strengthens the position of explainable AI (XAI) and hybrid approaches discussed in the previous paper [1].

8. Conclusion

This paper has updated the previous overview of AI tools in industrial control with three key dimensions of 2025. First, the choice between cloud and on-premise LLM deployment is not binary – the industrial standard has shifted towards hybrid architectures, where local models handle latency-sensitive tasks and data protection, while the cloud provides computational capacity for complex analyses. Second, building custom enterprise LLM systems is accessible even to medium-sized enterprises – the most effective path is a combination of RAG with fine-tuning of existing open-source models on enterprise metallurgical and operational datasets. Third, automated processing of technical documentation via RAG systems delivers demonstrable operational benefits: a 60–75% reduction in information retrieval time and a reduction in errors caused by outdated documentation.

For the steelmaking industry, these technologies – together with agentic AI systems – open new possibilities for heat optimisation, predictive maintenance and knowledge management. The key prerequisites for successful deployment remain a sound data infrastructure, compliance with the EU AI Act, and integration with human expertise in the spirit of Industry 5.0.

Hutnické listy č. 1/2026, roč. LXXIX
ISSN Online: 3029-8350

ACKNOWLEDGEMENT

This paper was produced within the research activities of VŠB – Technical University of Ostrava, Faculty of Materials Science and Technology.

„Originál tohoto článku byl publikován ve sborníku konference Oceláři 2026 a je zde uveřejněn se souhlasem autora.“

LITERATURE

[1] ŠPIČKOVÁ, Dagmar and Ivo ŠPIČKA. Artificial Intelligence Tools in Industrial Control. In: *Proceedings of the OCELÁRI 2025 Conference*. Rožnov pod Radhoštěm: TANGER, 2025.
[2] IOT ANALYTICS. *Industrial AI Market: 10 Insights on How AI is Transforming Manufacturing* [online]. 2024 [cit. 2026-02-20]. Available from: https://iot-analytics.com/industrial-ai-market-insights
[3] SOLO.IO. *What is Agentic AI? Definition, Benefits & Use Cases* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.solo.io/topics/ai-infrastructure/what-is-agentic-ai
[4] RASCHKA, Sebastian. *The State Of LLMs 2025: Progress, Problems, and Predictions* [online]. 2025 [cit. 2026-02-20]. Available from: https://magazine.sebastianraschka.com/p/state-of-llms-2025
[5] MIT SLOAN MANAGEMENT REVIEW; BOSTON CONSULTING GROUP. *Agentic AI, Explained* [online]. 2025 [cit. 2026-02-20]. Available from: https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained
[6] STRAITS RESEARCH. *Enterprise LLM Market Size, Industry Trends, Share and Report* [online]. 2025 [cit. 2026-02-20]. Available from: https://straitsresearch.com/report/enterprise-llm-market
[7] ALLGANIZE.AI. *Enterprise Guide: Choosing Between On-Premise and Cloud LLM and Agentic AI Deployment Models* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.allganize.ai/en/blog/enterprise-guide-choosing-between-on-premise-and-cloud-llm-and-agentic-ai-deployment-models
[8] S., Rosamma K. Small Language Models and Their Role in Hybrid AI Architectures for Big Data Analytics. In: *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*. 2024. Available from: https://doi.org/10.1109/icscna63714.2024.10863995
[9] UNIFIED AI HUB. *On-Prem LLMs vs Cloud APIs: When to Run Your AI Models On-Premise* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.unifiedaihub.com/blog/on-premise-llms-vs-cloud-apis-when-to-run-your-ai-models-on-premise
[10] F7I.AI. *LLM in Manufacturing: The 2025 Guide for Operations Leaders* [online]. 2025 [cit. 2026-02-20]. Available from: https://f7i.ai/blog/the-ultimate-guide-to-llms-in-manufacturing-from-co-pilot-to-competitive-edge-in-2025
[11] ARC ADVISORY GROUP. *Siemens Introduces AI Agents for Industrial Automation* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.arcweb.com/blog/siemens-introduces-ai-agents-industrial-automation
[12] CIO.COM. *Developing an Approach for Industry LLMs* [online]. 2025 [cit. 2026-02-20]. Available from: ttps://www.cio.com/article/4080214/developing-an-approach-for-industry-llms.html
[13] LEWIS, Patrick; PEREZ, Ethan; PIKTUS, Aleksandra et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Neural Information Processing Systems*. 2020.
[14] SUPERANNOTATE. *Fine-Tuning Large Language Models (LLMs) in 2025* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.superannotate.com/blog/llm-fine-tuning
[15] AGILESOFTLABS. *Build or Buy for Enterprise LLMs and When Custom Training Truly Matters* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.agilesoftlabs.com/blog/2025/12/build-or-buy-for-enterprise-llms-and
[16] OXMAINT. *From Technician Notes to Institutional Knowledge: LLM-Driven Knowledge Capture* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.oxmaint.com/blog/post/knowledge-capture-automation-technician-notes-llm
[17] OXMAINT. *From Technician Notes to Institutional Knowledge: LLM-Driven Knowledge Capture* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.oxmaint.com/blog/post/knowledge-capture-automation-technician-notes-llm
[18] SMS GROUP. *How AI is Transforming the Metals Industry* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.sms-group.com/insights/all-insights/how-ai-is-transforming-the-metals-industry
[19] MORDOR INTELLIGENCE. *Agentic AI in Manufacturing and Industrial Automation Market* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.mordorintelligence.com/industry-reports/agentic-artificial-intelligence-in-manufacturing-and-industrial-automation-market
[20] ABB. *Genix Copilot: Industrial AI Assistant* [online]. 2025 [cit. 2026-02-20]. Available from: https://search.abb.com/library/Download.aspx?DocumentID=9AKK108471A6627
[21] NATURE / SCIENTIFIC REPORTS. *Industrial Applications of Large Language Models* [online]. 2025 [cit. 2026-02-20]. Available from: https://www.nature.com/articles/s41598-025-98483-1