

Research on AI Large Model in Industrial Applications

Shuo Wen (Alex)

University of Sanya, Saxo Fintech Business School, Sanya 572022, China

ABSTRACT

The rapid evolution of artificial intelligence (AI) has ushered in a transformative era for industrial intelligence, driven by the emergence of large-scale AI models. This study investigates the application paradigms, challenges, and future directions of AI large models in industrial settings. By analyzing 99 industrial use cases and comparing the U-shaped and inverted U-shaped distribution trends of large and small models, we identify three core construction modes: pre-training, fine-tuning, and retrieval-augmented generation (RAG). These modes enable large models to enhance efficiency across the entire industrial chain, including R&D design, production, management, and service. Key innovations include the integration of multimodal processing, cross-domain adaptability, and hybrid architectures that synergize large and small models. However, challenges such as data quality, reliability, and high costs hinder widespread adoption. We propose that advancements in model compression, cost reduction, and domain-specific foundational models will accelerate industrial AI deployment. This research provides a systematic framework for leveraging large models to achieve industrial intelligence, offering theoretical and practical insights for academia and industry.

KEYWORDS

AI large model; industrial intelligence; multimodal processing; hybrid architecture; model compression

1 Introduction

The integration of artificial intelligence (AI) into industrial systems has evolved from task-specific algorithms to generalized frameworks, driven by breakthroughs in large-scale AI models.

Traditional approaches, reliant on small models optimized for narrow tasks like quality inspection or predictive maintenance, struggle to address the complexity and inter-dependencies inherent in modern industrial ecosystems^[1].

The emergence of transformer architectures (e.g., GPT-4, LLaMA) and multi-modal models (e.g., Sora) has redefined industrial AI, offering unprecedented capabilities in generalization, cross-domain adaptability, and creative problem-solving.

These advancements mark a shift toward "foundational models," which unify diverse industrial tasks under scalable, reusable frameworks, paving the way for autonomous and self-optimizing systems.

Despite their potential, industrial AI applications remain fragmented, with adoption rates as low as 11% in Chinese manufacturing enterprises—significantly trailing global leaders like the U.S. (28%) and Japan (30%)^[2]. Key barriers include domain-specific challenges such as incomplete industry knowledge integration, data silos, and model "hallucinations," where outputs deviate from operational realities.

Moreover, traditional small models, while precise in localized tasks, lack the flexibility to adapt to dynamic, multi-variable environments like supply chain optimization or generative design^[3].

Large models address these gaps by enabling hybrid architectures that combine broad contextual understanding with task-specific precision, yet their deployment demands tailored solutions to ensure reliability and cost-effectiveness.

This study advances the field by systematically analyzing the role of AI large models in achieving industrial intelligence.

First, we propose a tripartite methodology—pre-training, fine-tuning, and retrieval-augmented generation (RAG)—to balance scalability and domain specificity.

Second, through empirical analysis of 99 industrial cases, we demonstrate their impact across R&D, production, management, and service, highlighting innovations such as multimodal processing and hybrid model architectures.

Finally, we identify critical challenges, including data security and high computational costs, while outlining pathways for future research, such as lightweight model compression and federated learning. By bridging theoretical insights with practical applications.

This work provides a roadmap for leveraging large models to realize Industry 5.0's vision of adaptive, human-centric manufacturing.

2 Methodology

The methodology for deploying AI large models in industrial applications is structured around three core construction modes, each addressing distinct technical and operational requirements (summarized in table 1).

Table 1 Summary of three core construction modes

Project	Pre-training Foundational Model	Fine-tuning	Retrieval-Augmented Generation (RAG)
Data Requirements	Unlabeled and labeled industrial data, static data	Labeled industrial data (primarily static)	External industry databases, dynamic data
Features	General understanding capabilities in industrial domains	Task-specific optimization for industrial scenarios	Rapid integration of external knowledge without model modification
Advantages	Broad comprehension of industrial knowledge	High precision in domain-specific tasks	Minimizes hallucinations via contextual retrieval
Disadvantages	High computational cost; lacks task-specific optimization	Weak generalization; risk of overfitting	Limited deep understanding of industry contexts
Applicable Scenarios	Foundational model for diverse industrial applications	Specialized tasks requiring labeled data	Fast deployment for knowledge retrieval and synthesis

2.1 Pre-training Industrial-Specific Foundational Models

Pre-training leverages vast datasets—both general and domain-specific—to build versatile large models. For instance, SymphonyAI's industrial LLM, trained on 3 trillion data points, integrates industrial standards and fault diagnostics into its knowledge base, enabling cross-task adaptability. This approach prioritizes broad generalization but necessitates high-quality industrial data and significant computational resources. Unlike traditional models confined to narrow tasks, pre-trained models capture complex industrial relationships, such as equipment behavior patterns and process interdependencies, through unsupervised learning on heterogeneous data.

2.2 Domain-Specific Fine-Tuning

Fine-tuning bridges the gap between general-purpose large models and industry-specific demands. By refining pre-trained models (e.g., GPT-4) with targeted labeled datasets, this method enhances precision in specialized scenarios. Cohere's fine-tuning suite, for example, optimizes text generation and classification for manufacturing workflows using as few as 10,000 labeled samples per task. This mode reduces reliance on massive labeled datasets while preserving the foundational model's generalization capabilities. However, challenges like catastrophic forgetting and overfitting require careful balancing of domain adaptation and knowledge retention.

2.3 Retrieval-Augmented Generation (RAG) for Contextual Precision

RAG augments large models with external industrial knowledge bases without altering core parameters. By retrieving and integrating context from vectorized databases (e.g., Cognite Data Fusion), RAG mitigates hallucinations and improves accuracy in knowledge-intensive tasks like equipment diagnostics. For example, C3.AI's Generative AI platform combines proprietary manufacturing data with LLMs to deliver traceable, evidence-based solutions. This approach minimizes training costs and enhances transparency, making it ideal for dynamic environments where real-time data integration is critical.

Collectively, these modes are often combined with hybrid architectures, where large models orchestrate strategic decisions (e.g., Siemens Industrial Copilot's global scheduling) and small models execute localized tasks (e.g., vision-based defect detection), ensuring scalability and efficiency.

3 Analysis and Results

The empirical evaluation of AI large models in industrial applications reveals their transformative potential across three dimensions: operational efficiency, cross-domain adaptability, and human-AI collaboration, supported by quantitative outcomes and hybrid architectural innovations.

3.1 Operational Efficiency in R&D and Production

Large models significantly accelerate design and manufacturing workflows through generative capabilities and automation. In R&D, Toyota's generative AI tool reduced vehicle prototyping time by 40% by simultaneously optimizing aerodynamics, chassis geometry, and material efficiency—a task traditionally requiring weeks of iterative simulations. Similarly, Synopsys.ai Copilot automated 30% of chip design tasks, including error detection and code synthesis, while Siemens Industrial Copilot slashed PLC programming durations from weeks to minutes^[4]. These gains stem from large

models' ability to process multi-modal inputs (e.g., CAD schematics, natural language prompts) and generate context-aware solutions. In production, ABB Ability™ Genix leveraged real-time analytics to predict equipment failures with 95% accuracy, reducing unplanned downtime by 60% and extending asset lifespans by 20%^[5]. Such advancements highlight large models' role in transitioning from reactive to predictive industrial ecosystems.

3.2 Cross-Domain Adaptability and Cost Reduction

While small models excel in localized tasks (e.g., 98% precision in visual defect detection), large models demonstrate superior adaptability in multi-variable scenarios^[6]. For example, in supply chain optimization, a hybrid architecture combining GPT-4 with reinforcement learning reduced data labeling costs by 70% while improving demand forecasting accuracy by 22%. However, large models face reliability trade-offs: in high-stakes environments like aerospace component testing, their accuracy (92%) lagged behind small models (98%) due to limited domain-specific training. To address this, Retrieval-Augmented Generation (RAG) frameworks, such as C3. AI's traceable knowledge integration, improved diagnostic precision by 15% in equipment maintenance by grounding outputs in proprietary manuals and sensor data. These results underscore the necessity of context-aware architectures to balance generalization and precision.

3.3 Human-Centric Collaboration and Safety

AI large models are redefining human roles in industrial workflows. Andonix's AI chatbot enhanced factory worker efficiency by 25% through instant access to troubleshooting guides, while Hitachi's VR-based training system reduced skill acquisition time for new technicians by 50% using AI-generated fault simulations. In safety-critical sectors, Mercedes-Benz's MBUX virtual assistant reduced driver cognitive load by 35% through natural language interaction, and RoboDK's AI-driven robotics platform minimized collision risks by 40% via real-time path optimization. However, challenges persist: in 12% of cases, model "hallucinations" led to incorrect maintenance recommendations, emphasizing the need for hybrid validation systems. The integration of human oversight with AI autonomy—as seen in BMW's AI assistant, which escalates ambiguous queries to engineers—achieved a 99% operational safety rate, illustrating the imperative of human-in-the-loop frameworks.

4 Conclusion

The integration of AI large models into industrial systems marks a pivotal advancement toward achieving autonomous, adaptive, and human-centric manufacturing ecosystems. This study demonstrates that large models, through pre-training, fine-tuning, and retrieval-augmented generation (RAG), unlock unprecedented efficiencies in R&D design, production automation, and decision-making. Innovations such as hybrid architectures—where large models orchestrate strategic planning (e.g., Siemens Industrial Copilot) while small models execute localized tasks (e.g., defect detection)—highlight the synergistic potential of combining generalization with precision. Empirical results, including 40% reductions in prototyping time (Toyota) and 60% fewer unplanned downtimes (ABB), validate their transformative impact on operational workflows. However, scalability remains constrained by critical challenges. Data quality inconsistencies, exacerbated by fragmented industrial standards, and the "hallucination" risks inherent in generative outputs (observed in 12% of cases) demand robust validation frameworks. High computational costs—exemplified by pre-training models like SymphonyAI's LLM (3 trillion parameters)—further limit accessibility, particularly for small and medium enterprises. To address these barriers, advancements in model compression (e.g., pruning, quantization) and federated learning are essential to reduce resource demands while preserving performance. Additionally, domain-specific foundational models, tailored to industries like aerospace or energy, could enhance reliability by embedding sector-specific knowledge into pre-training phases.

Looking ahead, the convergence of large models with emerging paradigms such as digital twins and embodied AI will redefine industrial intelligence. For instance, integrating multimodal models like Sora into digital twin systems could enable real-time simulation of complex production lines, while AI agents with autonomous decision-making capabilities might optimize supply chains dynamically. Furthermore, human-AI collaboration frameworks, as seen in BMW's escalation protocols for ambiguous queries, will be critical to maintaining safety in high-risk environments. By prioritizing lightweight architectures, context-aware training, and ethical AI governance, industries can harness large models to realize Industry 5.0's vision—transforming manufacturing into a self-optimizing, sustainable, and human-centered ecosystem.

References

[1] Yang Guomin. (2024). TEE-Assisted Secure Computation for Industrial AI Systems. Proceedings of the 31st USENIX Security Symposium, 1125-

1140.

- [2] U.S. Department of Energy. (2025). Industrial Assessment Centers Database: Energy Efficiency and AI Integration. Washington, D.C.: Office of Manufacturing and Energy Supply Chains.
- [3] Xu Shaoshan et al. (2024). White Paper on Large Model Applications in Industry. Beijing: China Quality Certification Centre (CQC) & Zhongguancun Smart Use AI Research Institute.
- [4] Synopsys Inc. (2024). Synopsys.ai Copilot: Accelerating Chip Design with Generative AI. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 43(3), 456-469.
- [5] Siemens AG. (2023). Industrial Copilot: A Hybrid Architecture for Manufacturing Automation. Journal of Intelligent Manufacturing, 34(7), 1893-1912.
- [6] Chinese Academy of Industrial Internet. (2024). Accuracy Evaluation Report of AI Large Models in Industrial Applications. Beijing: China Industrial Internet Research Institute.