Journal of Emerging Intelligence and Engineering Technologies

Articles

Emerging Trends in Generative AI and Machine Learning for Enterprise Intelligence

Cheung Sha*

Faculty of Science and Technology, Hong Kong Institute of Technology, Hong Kong SAR China

Abstract

Generative artificial intelligence (AI), driven by advances in large language models (LLMs) and multimodal foundation models, has rapidly transitioned from a research curiosity to a central pillar of enterprise information technology strategy. Organizations across industries are integrating generative AI into knowledge management, customer engagement, software development, and operational decision-making, giving rise to the broader concept of enterprise intelligence: the systematic use of AI-driven insight generation, automation, and reasoning to enhance organizational performance. This paper presents a comprehensive review of emerging trends in generative AI and machine learning for enterprise intelligence, covering the evolution from traditional business intelligence and predictive analytics to generative and agentic AI paradigms, the architectural building blocks that enable enterprise-grade deployment — including retrieval-augmented generation (RAG), fine-tuning and parameter-efficient adaptation, and multi-agent orchestration — and representative enterprise application domains spanning knowledge management, customer service, software engineering, marketing, and enterprise decision support. The review further examines critical open challenges, including hallucination and factual reliability, data governance and privacy, integration with legacy enterprise systems, cost and computational overhead, and organizational change management. Finally, the paper outlines emerging research and practice directions, including agentic multi-step workflows, enterprise-grade retrieval and grounding architectures, small and domain-specialized language models, and governance frameworks for responsible enterprise AI adoption. This review is intended to serve as a consolidated reference for researchers and practitioners navigating the rapidly evolving landscape of generative AI-driven enterprise intelligence.

Keywords: Generative AI, Large Language Models, Enterprise Intelligence, Retrieval-Augmented Generation, Agentic AI, Business Intelligence, Fine-Tuning, AI Governance

Received: 28 July 2025
Revised: 10 August 2025
Accepted: 14 November 2025
Published: 14 December 2025

Introduction

Enterprises have long invested in data-driven decision-making, progressing through successive waves of technology: traditional business intelligence (BI) dashboards and reporting, statistical and machine learning-based predictive analytics, and, most recently, generative artificial intelligence (AI) built on large language models (LLMs) and multimodal foundation models. Unlike earlier analytical paradigms, which primarily summarized or predicted from structured historical data, generative AI systems can synthesize natural-language narratives, generate and review software code, draft documents, answer open-ended questions over unstructured enterprise knowledge, and increasingly act autonomously to complete multi-step tasks. This capability shift has catalyzed the emergence of the broader concept of enterprise intelligence: the systematic application of AI-driven insight generation, automation, and reasoning across an organization's knowledge, processes, and decision-making.

Industry surveys conducted through 2025 consistently report that a substantial majority of large enterprises have moved beyond initial pilots to production deployment of generative AI in at least one business function, with spending continuing to increase year over year. At the same time, the same surveys report persistent gaps between generative AI investment and measured return on investment, driven by challenges in data readiness, integration with legacy systems, and organizational change management, underscoring that technical capability alone does not guarantee enterprise value realization.

The technical foundation of enterprise generative AI has also evolved rapidly. Early enterprise deployments relied primarily on prompt engineering against general-purpose foundation models accessed via API. As enterprises sought greater factual reliability and integration with proprietary knowledge, retrieval-augmented generation (RAG) architectures, which ground model outputs in retrieved enterprise documents and data, became the dominant pattern for enterprise knowledge applications. More recently, agentic AI architectures, in which LLM-driven agents plan and execute multi-step tasks by invoking external tools, APIs, and enterprise systems, have emerged as a significant trend, extending generative AI from passive content generation toward active process automation.

This paper presents a comprehensive review of emerging trends in generative AI and machine learning for enterprise intelligence, with the following objectives: (1) to trace the evolution from traditional business intelligence to generative and agentic AI paradigms; (2) to systematically review the architectural building blocks enabling enterprise-grade generative AI deployment; (3) to survey representative enterprise application domains; (4) to critically examine open challenges; and (5) to identify promising future research and practice directions.

The remainder of this paper is organized as follows. Section 2 traces the evolution of enterprise intelligence. Section 3 reviews the architectural building blocks of enterprise generative AI. Section 4 surveys representative application domains. Section 5 presents a comparative analysis. Section 6 discusses open challenges. Section 7 outlines future directions, and Section 8 concludes the paper.

Evolution of Enterprise Intelligence

Traditional Business Intelligence and Predictive Analytics

Traditional business intelligence systems aggregate structured organizational data into dashboards, reports, and key performance indicators, primarily supporting retrospective and descriptive analysis. Predictive analytics, building on classical machine learning techniques, extended this capability to forecasting and risk scoring, enabling proactive rather than purely reactive decision support. Both paradigms, however, remain fundamentally constrained to structured, well-defined data schemas and require substantial manual effort to extend to new questions or unstructured data sources such as documents, emails, and support tickets, which constitute the majority of enterprise information.

The Generative AI Inflection

The emergence of large-scale transformer-based language models capable of general-purpose natural-language understanding and generation marked a qualitative shift in enterprise AI capability. Generative AI systems can interpret unstructured queries posed in natural language, synthesize information across heterogeneous and previously siloed data sources, and produce human-readable narrative outputs, code, or structured data, substantially lowering the technical barrier for non-specialist employees to extract value from organizational data and knowledge.

Toward Agentic Enterprise Intelligence

The current frontier of enterprise generative AI extends beyond content generation toward agentic systems capable of autonomously planning and executing multi-step workflows, invoking external tools and enterprise application programming interfaces (APIs), and coordinating across multiple specialized sub-agents. This trajectory reflects a broader shift from generative AI as a content-generation assistant toward generative AI as an active participant in enterprise business processes, with corresponding increases in both potential productivity gains and governance complexity.

Architectural Building Blocks for Enterprise Generative AI

Prompt Engineering and In-Context Learning

The simplest and most widely adopted pattern for enterprise generative AI deployment involves crafting effective prompts, including task instructions, few-shot examples, and output format specifications, to elicit desired behavior from a general-purpose foundation model without any model retraining. While fast and low-cost to implement, this approach is fundamentally limited by the model's context window and pretraining knowledge, and cannot incorporate proprietary or rapidly changing enterprise information beyond what is explicitly included in the prompt.

Retrieval-Augmented Generation (RAG)

Retrieval-augmented generation addresses the knowledge limitations of prompt-only approaches by retrieving relevant documents or data records from an enterprise knowledge base, typically indexed via dense vector embeddings in a vector database, and supplying the retrieved content as grounding context to the language model at inference time. This architecture has become the dominant pattern for enterprise knowledge management and question-answering applications, as it substantially reduces hallucination relative to ungrounded generation, keeps responses current with evolving enterprise data without requiring model retraining, and can provide source attribution to support verification. Recent developments in this space include agentic RAG, which incorporates iterative, multi-step retrieval and reasoning rather than a single retrieval pass, and hybrid retrieval strategies combining dense semantic search with traditional keyword-based methods to improve retrieval precision over diverse enterprise document types.

Fine-Tuning and Parameter-Efficient Adaptation

For applications requiring consistent domain-specific behavior, tone, or specialized task performance beyond what prompting or retrieval alone can achieve, enterprises increasingly employ fine-tuning, adapting a pretrained foundation model's parameters using domain-specific labeled or curated data. Parameter-efficient fine-tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) and adapter modules, enable this customization while updating only a small fraction of model parameters, substantially reducing the computational cost and infrastructure required relative to full model fine-tuning, and have become the preferred approach for most enterprise customization scenarios.

Agentic Multi-Step Workflows

Agentic AI architectures extend generative AI beyond single-turn content generation by enabling an LLM-driven agent to decompose a complex task into a sequence of steps, invoke external tools, APIs, or enterprise systems (such as customer relationship management or enterprise resource planning platforms) to gather information or take action, and iteratively refine its plan based on intermediate results. Multi-agent architectures further coordinate specialized sub-agents, each responsible for a distinct sub-task or domain, under an orchestrating controller. This paradigm enables automation of substantially more complex, end-to-end business processes than single-prompt generation, but introduces corresponding challenges in reliability, error propagation across steps, and auditability of autonomous decision chains.

Small and Domain-Specialized Language Models

In parallel with the growth of increasingly large general-purpose foundation models, enterprises have shown growing interest in small language models (SLMs) and domain-specialized models, either trained from scratch on curated domain data or distilled from larger models, to reduce inference cost, latency, and infrastructure requirements, particularly for high-volume or on-premise/ edge deployment scenarios where reliance on large third-party hosted models is impractical or undesirable for cost or data-sovereignty reasons.

Representative Enterprise Application Domains

Knowledge Management and Enterprise Search: RAG-based generative AI systems increasingly serve as intelligent interfaces over enterprise document repositories, wikis, and knowledge bases, enabling employees to pose natural-language questions and receive synthesized, source-attributed answers in place of traditional manual keyword search across disparate systems.

Customer Service and Engagement: Generative AI-powered conversational agents handle customer inquiries, draft support responses, and summarize customer interactions, with agentic architectures increasingly enabling more complex end-to-end resolution of customer requests, such as processing returns or updating account information, rather than purely informational responses.

Software Engineering: AI coding assistants integrated into development environments generate, explain, and review code, accelerating software development workflows; enterprise adoption has extended to automated test generation, code migration, and documentation generation, meaningfully altering developer productivity patterns.

Marketing and Content Operations: Generative AI supports drafting of marketing copy, personalized customer communications, and content localization at scale, though enterprises increasingly emphasize human review workflows to manage brand consistency and factual accuracy risks.

Enterprise Decision Support and Reporting: Generative AI is increasingly integrated into business intelligence tooling to enable natural-language querying of structured enterprise data and automated narrative generation summarizing analytical findings, lowering the barrier for non-technical stakeholders to interact with organizational data.

Comparative Analysis

Across the reviewed literature and industry reporting, several consistent patterns emerge in enterprise architectural choice. Retrieval-augmented generation has emerged as the default architecture for enterprise knowledge applications, favored for its ability to ground model outputs in current, proprietary data without the cost and latency of retraining, though overall system quality remains highly sensitive to retrieval pipeline design, including document chunking strategy, embedding model choice, and retrieval ranking. Fine-tuning and parameter-efficient adaptation are more selectively applied, typically reserved for applications requiring consistent domain-specific behavior at scale, where the upfront investment in curated training data and compute is justified by high query volume or stringent output consistency requirements.

Agentic multi-step workflows represent the most rapidly growing but least mature architectural pattern, with industry reporting through 2025 indicating substantial enterprise experimentation but comparatively fewer large-scale production deployments relative to RAG-based knowledge applications, reflecting ongoing concerns regarding reliability and governance of autonomous multi-step actions.

Small and domain-specialized models are increasingly adopted as a cost- and latency-optimization layer alongside, rather than as a wholesale replacement for, large general-purpose foundation models, with many enterprise architectures now routing queries across multiple models of varying scale based on task complexity.

Open Challenges

Hallucination and Factual Reliability: Generative models can produce plausible-sounding but factually incorrect or unsubstantiated output, a particularly consequential risk in enterprise contexts involving regulatory, financial, or customer-facing communications; while RAG architectures substantially mitigate this risk, they do not eliminate it, particularly when retrieval fails to surface relevant grounding context.

Data Governance and Privacy: Enterprise generative AI deployment requires careful governance of what proprietary or sensitive data is exposed to model providers, retrieval systems, and generated outputs, particularly for organizations operating under sector-specific regulatory regimes or data-residency requirements, motivating growing interest in on-premise and private-cloud deployment options.

Integration with Legacy Enterprise Systems: Realizing value from generative and agentic AI typically requires integration with existing enterprise systems, including customer relationship management, enterprise resource planning, and document management platforms, an integration effort that industry reporting consistently identifies as a leading barrier to moving from pilot to production deployment.

Cost and Computational Overhead: The inference cost of large foundation models, particularly for high-volume enterprise applications or agentic workflows involving multiple model calls per task, remains a significant and often underestimated component of total deployment cost, motivating interest in smaller specialized models and more efficient serving infrastructure.

Table 1: Comparative Summary of Enterprise Generative AI Architecture Patterns

Enterprise architecture pattern	*Core mechanism*	*Typical benefits*	*Typical limitations*
Prompt Engineering / In-Context Use	Task instructions and examples supplied directly in the model prompt	No training required, fast to deploy, low cost	Limited by context window, inconsistent on complex tasks
Retrieval-Augmented Generation (RAG)	Retrieves relevant enterprise documents/data to ground model responses	Reduces hallucination, keeps knowledge current, no retraining needed	Retrieval quality-dependent, added latency and infrastructure
Fine-Tuning / Parameter-Efficient Adaptation	Adapts a pretrained model's weights (fully or via LoRA/adapters) to domain data	Higher task accuracy, consistent domain-specific style/behavior	Requires labeled data, compute cost, risk of overfitting/drift
Agentic Multi-Step Workflows	LLM-driven agents plan, call tools/APIs, and execute multi-step tasks	Automates complex end-to-end processes, tool/system integration	Reliability and error-propagation risk, harder to audit/control
Small / Domain-Specialized Language Models	Compact models trained or distilled for a specific enterprise domain	Lower cost, faster inference, easier on-premise/edge deployment	Narrower general capability than large foundation models

Evaluation and Quality Assurance: Unlike traditional software, generative AI outputs are probabilistic and open-ended, complicating systematic testing and quality assurance; enterprises continue to develop evaluation frameworks and human-in-the-loop review processes to manage output quality at scale, an area that remains comparatively immature relative to the pace of deployment.
Organizational Change Management: Industry surveys consistently identify workforce adaptation, skills development, and organizational process redesign, rather than pure technical capability, as leading determinants of whether enterprise generative AI investments translate into measurable productivity or financial returns.

Future Directions

Maturation of Agentic Enterprise Workflows: Continued research and engineering investment into reliability, verifiability, and governance of multi-step agentic workflows is likely to determine the pace at which agentic AI moves from experimentation to widespread production deployment across enterprise processes.
Advanced Retrieval and Grounding Architectures: Further refinement of retrieval architectures, including iterative and agentic retrieval, structured knowledge-graph grounding, and hybrid retrieval strategies, is likely to continue improving the factual reliability and completeness of enterprise generative AI outputs.
Proliferation of Small and Domain-Specialized Models: Continued growth in capable small and domain-specialized language models is expected to expand enterprise deployment options for cost-sensitive, latency-sensitive, and data-sovereignty-constrained use cases, alongside continued use of large general-purpose models for more complex or open-ended tasks.
Standardized Governance and Evaluation Frameworks: The development of standardized enterprise AI governance frameworks, covering risk assessment, output evaluation, and regulatory compliance, is likely to become an increasingly central component of enterprise generative AI strategy as regulatory scrutiny of AI systems continues to increase globally.
Multimodal Enterprise Intelligence: Extension of generative AI capability beyond text to integrated reasoning over images, audio, video, and structured enterprise data represents a promising direction for broadening the scope of enterprise intelligence applications, particularly in domains such as manufacturing quality inspection, medical documentation, and multimedia content operations.

Conclusion

Generative AI and machine learning are reshaping enterprise intelligence, extending organizational data and decision-support capability from structured business intelligence and predictive analytics toward natural-language reasoning over unstructured knowledge and, increasingly, autonomous multi-step process automation. This review has traced this evolution, surveyed the core architectural building blocks of enterprise generative AI deployment — prompt engineering, retrieval-augmented generation, fine-tuning, agentic workflows, and domain-specialized models — and examined representative application domains spanning knowledge management, customer engagement, software engineering, marketing, and decision support. While enterprise adoption has accelerated rapidly, realizing consistent value remains contingent on addressing persistent challenges around factual reliability, data governance, legacy system integration, cost management, and organizational change. Continued progress in agentic workflow reliability, advanced retrieval and grounding architectures, and standardized governance frameworks is likely to define the next phase of generative AI-driven enterprise intelligence.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., "Language Models are Few-Shot Learners," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 1877-1901.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 9459-9474.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in Proc. Int. Conf. Learning Representations (ICLR), 2022.

[4] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," in Proc. Int. Conf. Learning Representations (ICLR), 2023.

[5] "An Automated Survey of Generative Artificial Intelligence: Large Language Models, Architectures, Protocols, and Applications," arXiv preprint, arXiv:2306.02781, 2024. Available: https://arxiv.org/html/2306.02781v4

[6] Deloitte, "State of Generative AI in the Enterprise, Quarter Four Report," Deloitte Insights, 2025. Available: https://www.markt-kom.com/wp-content/uploads/2025/01/Download-Deloitte-Report-State-of-Generative-AI-Englisch.pdf

[7] Deloitte, "The State of AI in the Enterprise," Deloitte Insights, 2026. Available: https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html

[8] Menlo Ventures, "2025: The State of Generative AI in the Enterprise," Menlo Ventures Perspectives, 2025. Available: https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/

[9] "Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review," Applied Sciences, vol. 16, no. 1, p. 368, 2025. Available: https://www.mdpi.com/2076-3417/16/1/368

[10] "Retrieval-Augmented Generation to Generate Knowledge Assets and Creation of Action Drivers," Applied Sciences, vol. 15, no. 11, p. 6247, 2025. Available: https://www.mdpi.com/2076-3417/15/11/6247

[11] "Agentic Retrieval-Augmented Generation for On-Premise Infrastructure: Enhancing Security and Performance for Enterprise Knowledge Management," IEEE Xplore, 2025. Available: https://ieeexplore.ieee.org/document/11436668/

[12] "From RAG to Context: A 2025 Year-End Review of RAG," RAGFlow, 2025. Available: https://ragflow.io/blog/rag-review-2025-from-rag-to-context

[13] Kong Inc., "72% Say Enterprise GenAI Spending Going Up in 2025, Study Finds," 2025. Available: https://konghq.com/blog/enterprise/enterprise-ai-spending-2025