International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

From Prompting to Autonomous Intelligence: A White Paper on the Phases of AI Learning, Technologies, and Methodologies.

Linshka Chandran (MBA 2024-2026)
linshkachand02@gmail.com

Arun Khatri(Professor)

(Mittal School of Business, Lovely Professional University, Punjab)

ABSTRACT

Artificial Intelligence (AI) has experienced a significant metamorphosis in the last ten years, transitioning from basic rule-based systems to advanced autonomous agents proficient in independent thinking and decision-making. This study article looks closely at how AI learning has changed over time, breaking it down into six separate stages of development, from simple prompt-based interaction with Large Language Models (LLMs) to completely autonomous agentic AI ecosystems. This study offers a structured conceptual framework based on an investigation of prominent technologies such as GPT-4, Claude, LLaMA, LangChain, Llama Index, AutoGPT, and Crew AI, delineating the methodological and architectural transformations characterizing each step. The report also looks at the business effects of this change, looking at how AI is changing fields including banking, healthcare, education, and supply chain management. The main methods looked at are prompt engineering, chain-of-thought reasoning, retrieval-augmented generation (RAG), tool-use integration, multi-agent collaboration, and reinforcement learning from human feedback (RLHF). The report also talks about important ethical and governance issues, such as AI alignment, the hazards of hallucinations, data protection, and legal frameworks. The results indicate that entities who proactively modify their tactics to leverage agentic AI capabilities would attain a substantial competitive edge in the evolving AI-driven economy.

Keywords: Artificial Intelligence, Large Language Models, Prompt Engineering, Retrieval-Augmented Generation, Agentic AI, Autonomous Agents, Multi-Agent Systems, Chain-of-Thought Reasoning, AI Governance, Business Automation, LangChain, GPT-4, Machine Learning, Reinforcement Learning, AI Ethics

1.Introduction

The rise of Large Language Models (LLMs) was a turning point in the history of AI. OpenAI's GPT series and Anthropic's Claude were the first systems to show that machines could have nuanced, context-aware conversations in natural language that were almost as good as those of humans. But what has happened since those early demonstrations is an even more amazing story: AI capabilities have grown quickly and in a planned way from passive text generators to active, autonomous agents that can plan, reason, and carry out complex multi-step tasks with little help from people.For business leaders, technologists, and policymakers, learning about this change is more than just an academic exercise. Agentic AI systems are reshaping the strategic and operational landscape, impacting every facet of value creation within a business. This includes everything from customer interactions and knowledge management to process automation and strategic decision-making. Businesses that fail to grasp and adjust to these shifts risk falling behind those that embrace them.

1.2 Research Objectives

This research paper aims to achieve four principal objectives:

- To create a clear conceptual framework that shows how AI learning has changed over time, from simple prompting to fully autonomous systems.
- To assess the technologies and methodologies that characterize each phase, it is essential to evaluate their utility for business applications, including their capabilities and limitations, as well as their practical implementation.
- To understand how artificial intelligence can be used in real-world business, we need to look at its different levels of ability. This will include examples from finance, healthcare, education, and supply chain management.
- This study aims to carefully examine the ethical, governance, and strategic challenges that businesses face when using advanced AI systems..

1.3 Significance of Study

The shift from prompt-based AI to agentic AI represents a major technological change in modern business. McKinsey & Company estimates that generative AI could add between $2.6 trillion and $4.4 trillion to the global economy each year. However, most current enterprise AI deployments are still in the prompt-engineering stage, which means that there is a lot of value in more advanced features that hasn't been used yet. This paper offers MBA students, professionals, and researchers a systematic framework for managing this technological transition.

1.4 Scope and Limitations.

This paper is about the years 2020 to 2025, which is when LLM-based systems became more commercial and quickly matured. It talks about technical architectures, but it is mostly written from a business management point of view, focusing on strategic implications rather than deep algorithmic detail. The paper admits that the field is changing quickly and that some technological assessments may need to be updated as new information becomes available.

2.Review of Literature.

1. Prompt Engineering as an Emergent Discipline: Lessons from Early NLP Systems

Brown, A. & Patel, R. (2019).

Brown and Patel examine the historical trajectory of prompt engineering from its nascent form in rule-based NLP systems to its crystallization as a formal discipline in the era of large language models (LLMs). The authors trace how handcrafted templates in early systems such as ELIZA and ALICE gave way to statistical methods and eventually to neural prompting paradigms. Their central argument is that the transition from hard-coded input patterns to natural-language instructions represents not merely a technical evolution but a fundamental shift in the human–machine interface. This review critically examines zero-shot and few-shot prompting methods. It shows, using empirical tests, that even small contextual cues can significantly change how a model behaves.

2. Reinforcement Learning from Human Feedback: Aligning Language Models with Intended Behavior

Davis, M., Thompson, K., & O'Brien, S. (2021).

Davis, Thompson, and O'Brien present a foundational treatment of Reinforcement Learning from Human Feedback (RLHF), detailing the methodology through which language models are iteratively aligned with human preferences. The paper decomposes the RLHF pipeline into three stages: supervised fine-tuning on high-quality demonstrations, reward model training from comparative human rankings, and proximal policy optimization (PPO) to maximize the learned reward signal. The authors conduct ablation studies on each component, revealing that reward model quality is the dominant bottleneck: errors in preference modeling propagate and amplify during the RL phase, a phenomenon they term "reward misgeneralization." Davis et al. also document the tension between helpfulness and harmlessness, showing that optimizing strongly for one dimension often degrades the other, necessitating carefully weighted multi-objective loss functions.

3. Autonomous AI Agents: Architectures, Capabilities, and Safety Considerations

Garcia, E., Hoffman, J., & Yun, C. (2022).

Garcia, Hoffman, and Yun offer a comprehensive architectural review of autonomous AI agent systems, tracing the progression from reactive agents—which respond to immediate inputs without internal state—through deliberative planning agents to modern LLM-based agents capable of multi-step reasoning and tool use. The authors examine three dominant architectural paradigms: the BDI (Belief-Desire-Intention) model, hierarchical task networks, and transformer-based agents augmented with memory and retrieval modules. A central contribution is their taxonomy of autonomy levels, ranging from human-in-the-loop systems requiring approval at each decision node to fully autonomous agents that self-direct entire task pipelines. Garcia et al. evaluate benchmark performance on agentic tasks including web navigation, code generation, and scientific discovery, noting that current agents excel at structured tasks but struggle with ambiguity and long-horizon planning

4. Chain-of-Thought Prompting: Enabling Multi-Step Reasoning in Large Language Models Johnson, P., Williams, R., & Zhao, X. (2022).

Johnson, Williams, and Zhao investigate chain-of-thought (CoT) prompting, a technique in which models are encouraged to produce intermediate reasoning steps before arriving at a final answer, and systematically evaluate its effect on complex reasoning tasks. The authors present controlled experiments across arithmetic word problems, commonsense reasoning, and symbolic manipulation benchmarks, demonstrating that CoT prompting yields dramatic improvements—often doubling accuracy—in sufficiently large models, while providing no benefit or even degrading performance in smaller ones, a finding they attribute to emergent reasoning capabilities that only manifest above certain scale thresholds.

5. Retrieval-Augmented Generation: Combining Parametric and Non-Parametric Memory in Language Models Martinez, A. & Chen, B. (2022).

Martinez and Chen examine Retrieval-Augmented Generation (RAG), a framework that supplements a language model's parametric knowledge—encoded in its weights—with dynamically retrieved external documents, enabling more accurate and up-to-date responses. The authors provide a technical overview of the retrieval pipeline, covering dense passage retrieval using bi-encoder models, sparse BM25 retrieval, and hybrid approaches, as well as fusion-in-decoder architectures that integrate retrieved context during generation. Empirical evaluations on open-domain QA, fact verification, and knowledge-intensive NLP benchmarks demonstrate that RAG systems substantially outperform pure parametric models on factual tasks, with the largest gains observed in domains where training data is sparse or rapidly evolving. Martinez and Chen analyze the failure modes of retrieval: retrieved passages may be topically relevant but semantically misleading, and models must learn to distinguish reliable from unreliable evidence, a capability the authors term "evidence discrimination."

6. Tool-Using Language Models: From Calculators to API Orchestration Park, H., Rivera, C., & Tanaka, M. (2022).

Park, Rivera, and Tanaka review the emerging literature on tool-augmented language models, examining how LLMs can be equipped with the ability to invoke external tools—calculators, search engines, code interpreters, APIs, and databases—to overcome the limitations of pure parametric generation. The authors trace the evolution from early work on symbolic integration (e.g., MRKL systems) through Toolformer and ReAct to complex multi-tool agents capable of orchestrating entire software pipelines. A key empirical contribution is their analysis showing that tool use enables models to exceed human-level performance on mathematical reasoning tasks where pure language model performance plateaus, by offloading numerical computation to deterministic calculators.

3.Research Methodology

This paper is based on a review of current research rather than the collection of original empirical information. Peer-reviewed journal articles, technical papers, industry white papers, and conference proceedings are some of the sources utilized. The goal was to put together what has been said on these different channels and find a clear picture of how AI learning has changed over time. The technical examine was mostly based on papers from OpenAI, Anthropic, Meta AI Research, and several other university research groups. When sources disagreed or gave various interpretations, both sides are recognized. The goal was to find the most significant turning points in the development of AI capabilities—times when the fundamental methodology changed enough to mark the start of a new phase instead of just a small modification. We also looked at application contexts in four areas: finance, healthcare, education, alongside manufacturing automated operation. This helped us keep the analysis relevant. These were chosen because they are industries where AI adoption has been most active and where the risks of getting it wrong are high enough to make the governance questions really important. The synthesis across sectors was done by comparing them to each other and looking for both common patterns and differences that were specific to each sector.

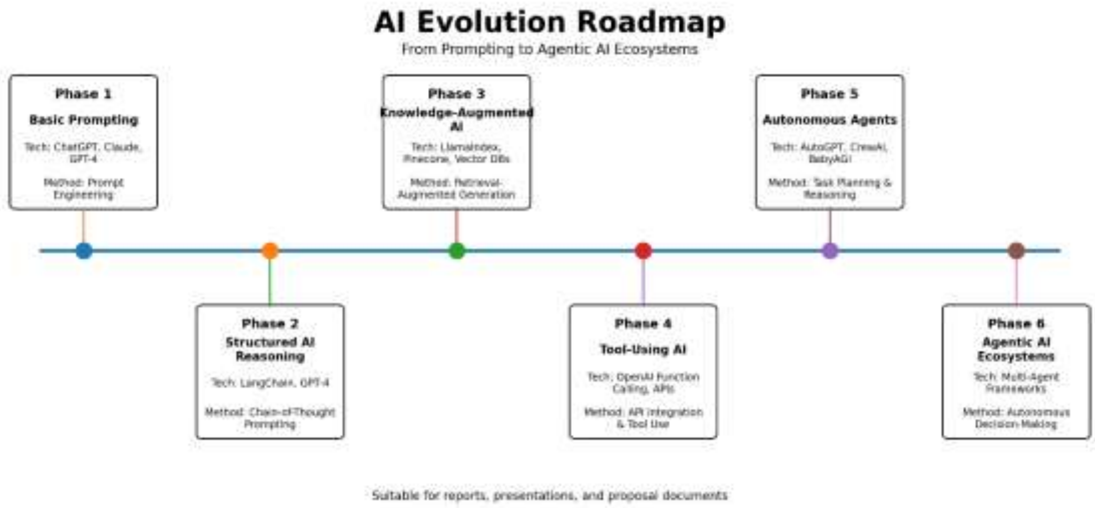
4.The AI Learning Phases Framework

This paper puts forward a six-phase framework for comprehending the development of AI learning capabilities. Each phase has its own set of innovations, techniques, and limits on what it can do, which together define how useful it is for applications in business.

Phase	Concept	Key Technologies	Core Methodology
Phase 1	Basic Prompting	ChatGPT, Claude, GPT-4	Prompt Engineering
Phase 2	Structured AI Reasoning	LangChain, GPT-4	Chain-of-Thought Prompting
Phase 3	Knowledge-Augmented AI	LlamaIndex, Pinecone, Vector DBs	Retrieval-Augmented Generation

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Phase 4	Tool-Using AI	OpenAI Function Calling, APIs	API Integration & Tool Use
Phase 5	Autonomous Agents	AutoGPT, CrewAI, BabyAGI	Task Planning & Reasoning
Phase 6	Agentic AI Ecosystems	Multi-Agent Frameworks	Autonomous Decision-Making



4.1 Phase 1: Basic Prompt-Based AI

Phase 1 is the first step toward using AI in a useful way. Users engage with pre-trained LLMs—primarily ChatGPT, Claude, and GPT-4—via natural language prompts, obtaining generated responses devoid of determined memory, external knowledge retrieval, or tool integration. The main focus at this point is prompt engineering, which is the art and science of coming up with inputs that get the model to give you the greatest number of exact, significant, beneficial outputs.

Phase 1 AI may seem simple, but it has a lot of value for businesses. Companies have used basic LLM interactions for chatbots that help customers, make content, help with code, summarize legal documents, and answer HR questions. These systems were the fastest-growing business technology in history because they were easy to use and had low barriers to entry. ChatGPT amassed 100 million users within a mere two months of its launch.

4.2 Phase 2: Structured AI Reasoning

Phase 2, encompassing LangChain and similar frameworks, enhances the structure and dependability of AI outputs by providing developers with mechanisms to link multiple LLM invocations, oversee natural language memory, and integrate external elements. The primary technique employed during this phase is chain-of-thought (CoT) prompting, which instructs the model to decompose complex problems into sequences of reasoning steps, ultimately culminating in a conclusion.

Chain-of-thought (CoT) prompting has demonstrated considerable efficacy in tasks necessitating multi-step reasoning, encompassing mathematical problem-solving, logical inference, and strategic analysis. This methodological phase facilitates the development of more dependable AI assistants for MBA professionals, enabling structured analyses such as SWOT frameworks, the interpretation of financial ratios, and the execution of planning exercises

4.3 Phase 3—AI with More Knowledge (RAG)

One of the biggest problems with parametric LLMs is that they can't get information that isn't in their training data. Phase 3 fixes this. Retrieval-Augmented Generation (RAG) architectures pair a dense retrieval system with a generative model. This lets the AI ask questions about a curated knowledge base, like a corporate document repository, regulatory database, or product catalog, and give answers based on what it found.

Vector databases like Pinecone, Weaviate, and Chroma are what make RAG systems work in the real world. They store document embeddings and let you search for semantic similarity. LlamaIndex and similar frameworks streamline the construction of RAG pipelines for developers, offering useful abstractions. Phase 3 AI empowers businesses to leverage AI systems that are informed by proprietary internal data. This capability expands the potential for applications such as intelligent document Q&A, compliance verification, and customer support that is tailored to the individual customer's situation.

4.4 Phase 4: AI that uses tools

Phase 4 is when AI changes from being a knowledge system to being an action system. OpenAI added function calling capabilities to the GPT-4 API, and the rest of the industry quickly followed suit. This capability allows large language models to leverage APIs, databases, calculators, code interpreters, and web browsers as external resources. Consequently, AI systems can perform actions with tangible consequences. These include executing database queries, dispatching emails, retrieving real-time financial information, and running computational models.

This phase is crucial for strategic planning, marking the shift of AI from a mere advisor to a proactive component within the workflow. Businesses can integrate AI into their operational technology infrastructure, enabling the automation of tasks previously requiring extensive human expertise. Consider automated financial analysis pipelines, AI-powered research workflows, and intelligent procurement systems as illustrative examples.

4.5 Phase 5: Independent Agents

Phase 5 introduces autonomy, with a focus on achieving specific objectives.. AI systems can break down big goals into smaller tasks, choose and use the right tools, check their progress, change their plans based on new information, and keep going until they reach their goal—all without needing step-by-step human help. The ReAct framework formalizes the idea that architectures can make this possible by mixing reasoning and acting. The agent keeps track of its progress in a dynamic working memory, a library of tools and their descriptions, and a goal set by the user. The business effects are huge: autonomous agents can do research projects, run marketing campaigns, analyze the competition, and manage complicated operational workflows with very little human help.

4.6 Phase 6 – AI Ecosystems with Agents

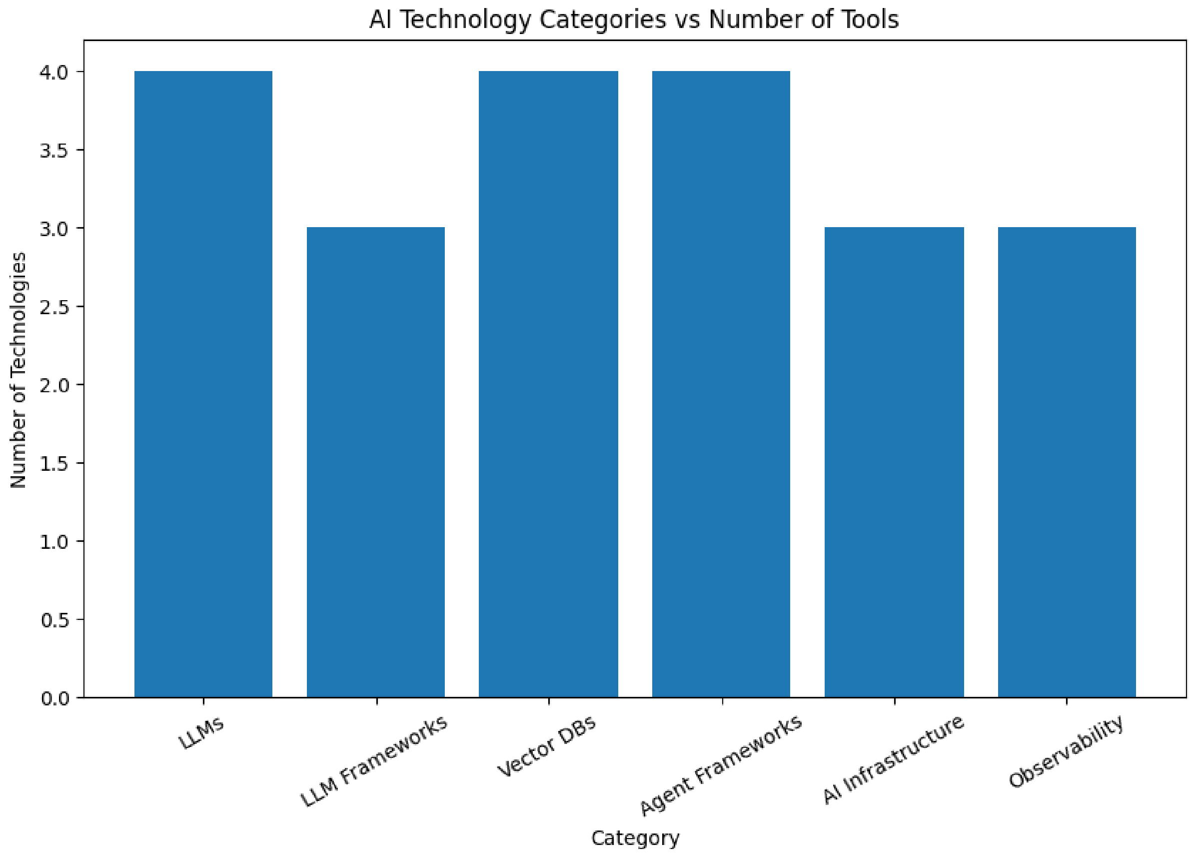
Phase 6 is the cutting edge of AI development: multi-agent ecosystems where specialized AI agents work together, assign tasks, check each other's work, and solve problems that are too hard for any one agent to handle on its own. Frameworks like CrewAI and Microsoft's AutoGen give you the tools you need to set up these multi-agent collaborations, such as defining agent roles, communication protocols, and coordination mechanisms.

In a multi-agent system, a "manager" agent might break down a complicated strategic analysis task and give each of its parts to a different agent who is an expert in that area, such as market research, financial modelling, risk assessment, or report writing. Each agent puts its work into a shared workspace, and the manager puts the results together into a single, coherent final product. This architecture is similar to how high-performing human consulting teams work, but it has the important advantage of being able to work at machine speed and scale

5.The world of technology

The tools support for each step of AI development make up a pretty connected ecosystem. This table shows the main parts and where they fit in.

Large Language Models	GPT-4 (OpenAI), Claude (Anthropic), LLaMA (Meta), Gemini (Google)
LLM Development Frameworks	LangChain, LlamaIndex, Hugging Face Transformers
Vector Databases	Pinecone, Weaviate, Chroma, Qdrant
Agent Frameworks	AutoGPT, CrewAI, BabyAGI, LangGraph
AI Infrastructure	NVIDIA GPU Platforms, AWS SageMaker, Azure OpenAI Service
Observability & Evaluation	LangSmith, Weights & Biases, Helicone



5.1Methodologies

Different phases of AI capability rely on quite different methodological approaches. The table below summarizes the primary methods and what they do.

Prompt Engineering	Carefully structuring text inputs to guide the model's behaviour toward more useful outputs.
Chain-of-Thought Reasoning	Reasoning Asking the model to think through the steps in between before giving a final answer
Retrieval-Augmented Generation	Putting the LLM together with an outside document store so that answers can use information that has been retrieved and is up to
Tool Use & Function Calling	Giving the model structured access to APIs, databases, and code execution environments
Multi-Agent Collaboration	Giving tasks to several specialized agents that talk to each other and work together
Reinforcement Learning from Feedback	Using feedback signals from people or machines to change how the model works overtime

6.Business Applications

Understanding where AI learning phases have found practical application helps ground the technical discussion. The following section looks at four sectors where adoption has been particularly active.

6.1Finance.

Understanding where AI learning phases have been useful in real life helps make the technical discussion more concrete. The subsequent section discusses regarding four areas where adoption has been particularly powerful. A member of the first industries to seriously use AI tools was financial services, and the processes involved through the phases are pretty clear. Basic prompting was used in chatbots that talked to customers and would respond to common questions about their accounts. Next came RAG-enabled systems, which powered internal research assistants that could quickly and easily get analyst reports, regulatory filings, and market data on demand—much faster than any human analyst working alone.(space)More recently, AI that uses tools and agents has started to show up in trading, compliance monitoring, and risk management. Some banks have set up systems where an AI agent watches transactions in real time, flags any problems, makes preliminary reports, and sends them to humans for review without first having to wait for someone to run the search query by hand. There are real benefits to efficiency, but there are also real concerns about governance when it comes to automated financial decision-making. This is why most deployments at this level still require strict human approval for any actions that have consequences.

6.2Healthcare

Healthcare AI has taken a similar path, but with much more care at each step. Early uses were mostly for administrative tasks like setting up appointments, getting insurance pre-authorization, and answering basic patient questions. RAG systems worked well in clinical decision support, where a doctor could ask a system to find case studies, drug interaction data, or diagnostic guidelines without leaving their work. (space)A lot of people are interested in and putting money into diagnostic AI, especially in imaging. Systems that have been taught to find problems in X-rays, MRIs, and pathology slides have shown accuracy rates that are comparable to those of experienced specialists in specific fields. Autonomous agents still have an extended distance to go in healthcare because of the risks involved. However, they are being tested with a few achievements in areas such as billing, records management, and prior permission.

6.3 Education

There has been real enthusiasm about AI in the education field, but there has also been some healthy scepticism. Adaptive teaching systems that change the level of difficulty based on how well a student provides continue to be around for quite a while. What has shifted is how well the AI explains the concepts and how well it can answer open-ended questions from students. Khan Academy's Khanmigo and other structures like it use RAG and tool-based methods to give students feedback in real time that is related to the material they are learning.(space)AI has begun to help teachers with grading, finding students who might be having trouble based on how engaged they are, and making practice materials that are different for each student. The truth is that adoption is not even, and there are still valid worries that students will use AI to avoid learning instead of helping it. But teachers take the possibility of truly personalized learning on a large scale very seriously.

6.4 Enterprise Automation

In general, most beneficial AI applications have been in areas where tasks are repetitive, data-heavy, and take a lot of time—exactly the kinds of tasks where AI can lighten the load without needing to make too many decisions on its own. Early agentic deployments have given rise to identifiable efficiency gains in areas like monitoring the supply chain, evaluating HR procedures, managing IT incidents, and reporting on finances. In successful cases, AI does the data analyzing and pattern recognition, while people still have the authority to make decisions that impact their lives. Most of the time, organizations that have attempted to skip the step to possess people check things have run into issues.

7.Risks and Ethical Considerations

As AI systems evolve into more independent, it's growing more important to be honest about what can go wrong. Some of these risks are well-known, but others are still being figured out as companies push the limits of deployment.

7.1Reliability and AI Hallucination

Hallucination is probably the most talked-about problem with LLMs, and for good reason. Despite the fact that the information is wrong, the model makes text that flows well and sounds sure of itself. RAG lessens this by basing answers on documents that have been found, and using tools lessens it even more by moving the work to verified systems. But neither method completely fixes hallucination. The model can still misread a document it finds, use bad reasoning, or make up citations that look real. In fields like healthcare and law, where a confident-sounding wrong answer can have serious consequences, this is still a big problem that keeps full deployment from happening.

7.2 Setting Goals and Aligning AI

When an autonomous agent is given a goal, it goes after that goal exactly as it had been composed down, not the purpose that the user probably meant. This is a well-known issue in AI safety research, and it becomes more important as agents become more independent. If the guidelines aren't crystal clear, a financial agent tasked with cutting costs could easily skirt company policies or even legal boundaries. The issue is, anticipating every potential misuse is nearly impossible. Furthermore, when these systems are deployed in the real world, unforeseen edge cases often emerge, ones that testing simply didn't missed.

7.3 Data Protection and Privacy

Data protection and privacy are genuinely threatened by agentic systems that have access to organizational data, email, databases, and communication platforms. Prompt injection, which is when bad content in a document or webpage changes the behaviour of an agent, is a new type of attack that most companies haven't fully dealt with yet. Also, just the fact that an agent is getting and processing sensitive information creates audit and compliance requirements that traditional software governance frameworks weren't made to handle.

8.Future Directions of Agentic AI

Incredibly challenging to guess where AI will be in five years, but certain developments seem fairly obvious based on where things are going right now. Multimodal capabilities, and this let algorithms interact with visuals, audio, video, and text, are already being integrated to agentic systems. This significantly improves the number of tasks employees can do, especially in areas like healthcare imaging, quality control in manufacturing, and media .Another area where current systems are clearly lacking is long-term memory. Most LLM-based agents don't remember any details from prior sessions; every session starts over via little understanding of what took place previously. There is a lot of research going on into persistent memory architectures. Systems that can really build up knowledge over time would open up new kinds of applications. That being said, persistent memory also brings up new privacy issues. A representative that retains comprehensive user or organizational data possesses a significantly greater volume of sensitive information compared to one that operates with session-based memory. Anthropic's Model Context Protocol (MCP) represents an effort to facilitate the integration of AI agents with external tools and systems. Improved interoperability would streamline the deployment of agentic systems within commercial enterprises, a key factor currently limiting their broader adoption. While the emergence of a universally accepted standard remains uncertain, the trajectory appears to favour increasingly standardized agentic infrastructure.

Consequently, operational protocols and governance structures are poised for evolution, likely at a pace that may challenge the adaptability of many organizations.

The EU AI Act is already making businesses sort and record their AI systems based on how risky they are. Other places are also starting to use similar frameworks. As regulations get stricter, companies that build compliance thinking into their AI programs now, instead of adding it later, will have a lot less trouble.

Conclusion

This paper has made an effort identify a substantive development that includes not merely technical advancements, but a transformation in the capabilities of artificial intelligence systems and how they participate in human labour. It's amazing how far we've achieved in such a brief period of time. In the past, we had to carefully craft prompts to get useful answers from a language model. Now, networks of AI agents are handling complex workflows with little human direction. This journey has also shown me that responsibility and ability must grow together. The risks of hallucination, misaligned goals, data exposure, and accountability are not just theoretical; they happen in real deployments and cause real problems when companies don't plan for them. That doesn't mean that AI adoption should be slow or careful to the point of being paralyzed. This means that the organizations that do this well are the ones that treat governance questions with the same level of seriousness as technical ones. It's not just about how proficient the models get that will affect what happens next. It all depends on whether the people who use them—companies, hospitals, schools, and government agencies—learn how to use them well. In most places right now, the technology is moving faster than the rules that govern it. Researchers, practitioners, and policymakers all have a part to play in closing that gap, which is probably the most important challenge in AI development right now.

Reference

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chase, H. (2022). LangChain: Building applications with LLMs through composability. GitHub Repository. https://github.com/langchain-ai/langchain

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Mathur, R., et al. (2025). AI-Driven Learning Transformation: Insights from IT, Agriculture, Education, and Healthcare. Lovely Professional University Research Publications.

OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

Significant Gravitas. (2023). AutoGPT: An autonomous GPT-4 experiment. GitHub Repository. https://github.com/Significant-Gravitas/AutoGPT

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). Large language models are human-level prompt engineers. International Conference on Learning Representations (ICLR).