Preprints.org

Article

Not peer-reviewed version

AI for Productivity in the Age of Agentic AI

Zhiheng Xi *, Enyu Zhou *, Xinyu Fang *, Zhe Sun, Baodai Huang, Jiajun Sun, Bicheng Deng, Zhihao Zhang, Wenxiang Chen, Jiazheng Zhang, Shichun Liu, Xin Guo, Zhikai Lei, Junke Wang, Senjie Jin, Yang Nan, Yajie Yang, Rui Zheng, Hang Yan, Yuchen Tian, Mengyue Yang, Yinghui He, Fang Wu, Cheng Qian, Xuandong Zhao, Yingcheng Wu, Mingchen Zhuge, Zhenfei Yin, Ling Yang, Jun Wang, Kejun Ying, Philip Torr, Tao Gui, Zuxuan Wu, Xipeng Qiu, Yu-Gang Jiang, Qi Zhang, Xuanjing Huang

Posted Date: 18 August 2026

doi: 10.20944/preprints202608.1248.v1

Keywords: agentic AI; AI for productivity

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

AI for Productivity in the Age of Agentic AI

Zhiheng Xi [1,*], Enyu Zhou [1,*], Xinyu Fang [2,*], Zhe Sun [1], Baodai Huang [1], Jiajun Sun [1], Bicheng Deng [1], Zhihao Zhang [1], Wenxiang Chen [1], Jiazheng Zhang [1], Shichun Liu [1], Xin Guo [1], Zhikai Lei [3], Junke Wang [1], Senjie Jin [1], Yang Nan [1], Yajie Yang [1], Rui Zheng [1], Hang Yan [1], Yuchen Tian [4], Mengyue Yang [5], Yinghui He [6], Fang Wu [7], Cheng Qian [8], Xuandong Zhao [9], Yingcheng Wu [7], Mingchen Zhuge [10], Zhenfei Yin [11], Ling Yang [6], Jun Wang [12], Kejun Ying [7], Philip Torr [11], Tao Gui [1], Zuxuan Wu [1], Xipeng Qiu [1], Yu-Gang Jiang [1], Qi Zhang [1] and Xuanjing Huang [1]

[1] Fudan University, China
[2] Zhejiang University, China
[3] East China Normal University, China
[4] Hong Kong Baptist University, China
[5] University of Bristol, UK
[6] Princeton University, USA
[7] Stanford University, USA
[8] University of Illinois Urbana-Champaign, USA
[9] University of California, Berkeley, USA
[10] King Abdullah University of Science and Technology, China
[11] University of Oxford, UK
[12] University College London, UK
* Correspondence: zhxi22@m.fudan.edu.cn (Z.X.); enyuzhou.zoey@gmail.com (E.Z.); fangxinyutju202009@126.com (X.F.)

Abstract

Throughout history, transformative technologies have repeatedly reshaped production systems and economic activity. In the current wave of AI development, Agentic AI systems are emerging as a new class of productive instruments capable of goal perception, multi-step planning, environmental interaction, and autonomous workflow coordination. Beyond automating routine tasks, these systems have the potential to augment knowledge work, reorganize decision-making, and reshape the relationship between human labor, organizational routines, and machine intelligence. This development gives rise to the emerging research direction of AI for Productivity (AI4Productivity), which examines how advanced AI systems affect productivity across economic and social contexts. In this paper, we conducts an in-depth investigation for explaining how Agentic AI creates productivity. We define AI4Productivity as the use of Agentic AI in real-world workflows to deliver economically valuable work with less human effort, and identify two core mechanisms of productivity gains: *Efficiency*, which reduces the time and labor required for existing tasks, and *Expansion*, which makes previously infeasible work economically viable. We further characterize the evolution of AI systems through four stages: conversational assistant, reactive operator, adaptive coordinator, and self-governing production system. Building on this framework, we review representative AI productivity tools and examine the sectoral diffusion of Agentic AI across three levels of adoption: early-stage exploration in manufacturing, agriculture, real estate, and government services; broader integration in trade, media, education, and legal services; deeper workflow orchestration in information technology, finance, and healthcare. Finally, we discuss key challenges including safety, long-horizon reliability, academia-industry gaps, adoption barriers, governance, and labor-market uncertainty. This paper provides a structured account of how Agentic AI is reshaping productivity and outlines future directions for reliable and socially beneficial AI-driven production systems. We hope this paper will serve as a foundation for advancing AI4Productivity research and guiding the design of AI systems that deliver reliable, scalable, and broadly beneficial productivity gains.

Keywords: agentic AI; AI for productivity

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

1. Introduction

Across diverse traditions of economic thought—from classical political economy [1] and Marxian political economy [2] to modern innovation economics [3]—technological breakthroughs have long been recognized as defining the boundaries of economic eras. From the steam engine of the Industrial Revolution, to electrification enabling mass production, to computers extending the limits of cognitive labor, each technological leap has redefined the frontier of human capability.

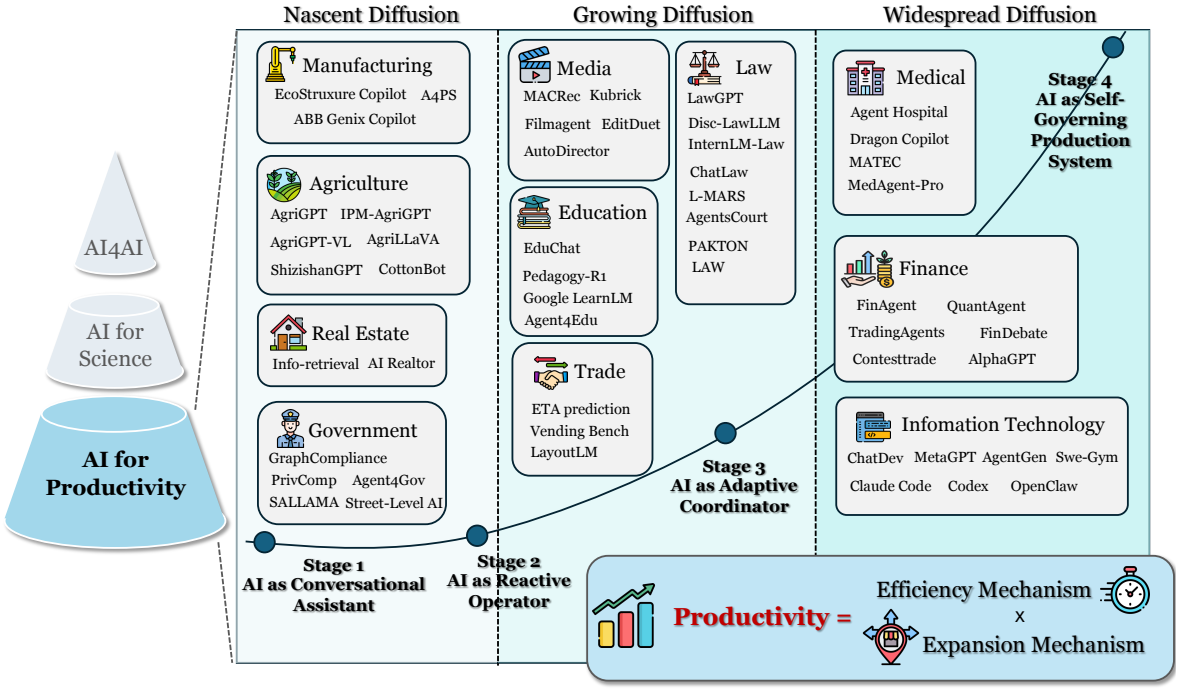


Figure 1. Overview of how agentic AI reshapes productivity through two mechanisms and four stages of delegated production responsibility. Using this framework, we assess how agentic AI is currently embedded across nine industries and extract cross-sector insights about where productivity gains are already visible, where they remain constrained, and what conditions are required for deeper diffusion.

"What distinguishes the different economic epochs is not what is produced, but how it is produced, by what instruments of labour."

— Karl Marx, Capital: A Critique of Political Economy

With the evolution of Large Language Models (LLMs), AI agents are demonstrating a growing capacity to execute complex tasks and deeply engage with the real world [4–9]. Emerging AI agent systems such as `OpenClaw` [10] and `Claude Code` [11] already illustrate their growing impact on practical workflows and developer ecosystems. This shift not only introduces a new technological paradigm, but also establishes agents as a new class of instruments of production. Compared with tools of previous eras, AI agents exhibit two fundamentally transformative properties: **generality** and **autonomy**. These enable them to operate across task boundaries and complete an autonomous closed loop—from environment perception and goal decomposition to decision-making and execution—without continuous human intervention [12,13], unlike traditional machines and software constrained by predefined rules and domain-specific programming [14].

Agentic AI should therefore be understood as a new productive instrument. Its importance lies not only in improving isolated task performance, but in changing how work is executed, coordinated, and delegated. The central question of this paper is not whether agentic AI can improve isolated task performance, but under what conditions technical capability is converted into sustained productivity gains.

We argue that this conversion depends on two mechanisms and one delegation process. The **efficiency mechanism** reduces the human effort required for existing output by automating parts of

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

complex workflows. The **expansion mechanism** makes valuable work feasible that would otherwise remain too costly, slow, or complex. These mechanisms become economically consequential only when production responsibility is gradually delegated from humans to AI systems. We therefore identify four stages of delegation: conversational assistant, reactive operator, adaptive coordinator, and self-governing production system.

Using this framework, we assess how agentic AI is currently embedded across nine industries and extract cross-sector insights about where productivity gains are already visible, where they remain constrained, and what conditions are required for deeper diffusion. Our contribution is not only to map deployed systems and benchmarks, but to explain why productivity gains are uneven across sectors. The evidence suggests that deeper diffusion depends on the digitization of production objects, feedback verifiability, error tolerance, workflow modularity, and accountability structures.

The remainder of the paper advances this argument as follows.

- **Section 2** establishes the conceptual lens by defining productivity as the outcome of how human labor is reorganized through productive instruments.
- **Section 3** defines the productivity mechanisms and delegation stages that connect technical capability to workflow-level productivity.
- **Section 4** uses deployed systems and benchmarks as evidence for how agentic AI begins to convert capability into economic work.
- **Sections 5–7** compare diffusion patterns across industries, distinguishing bounded assistance, bounded execution, and scaled orchestration.
- **Section 8** identifies structural constraints that explain why productivity diffusion remains uneven.
- **Section 9** synthesizes the implications for durable real-world value and delegated agency.

2. Conceptual Lens: Productivity and Productive Instruments

2.1. Conceptual Background: Productivity and Productive Instruments

This section introduces the conceptual lens used throughout the paper: productivity as the outcome of how human labor is reorganized through productive instruments. We draw on foundational concepts from economics and the history of technology not to advocate any particular political or ideological position, but to distill the notions of productivity and productive instruments into a structured lens for analyzing the economic impact of agentic AI. Productivity is commonly measured as output per unit of labor input, such as output per hour worked or GDP per hour worked [15]. This definition is useful for empirical analysis, but it does not explain the mechanisms through which major productivity gains are generated. Economic growth theory has long emphasized technological change as a central source of sustained productivity improvement [16]. To understand where productivity growth comes from, we must therefore look not only at labor input, but also at the technologies and other non-human resources that shape the production process.

The concept of productive instruments provides a critical lens for understanding these mechanisms. In the classical political economy, production depends on the combination of human labor and the means of production [2]. Within this broader category, productive instruments include the tools, machines, and technical infrastructures that directly shape the way work is carried out, coordinated, and scaled [17]. From this perspective, productivity depends not only on labor effort, but also on how labor is organized through these instruments. Therefore, changes in productive instruments often lead to broader changes in productivity.

In line with this framework, the major technological changes coincide with broader economic transformations [18,19]. Throughout the history of industrial development, breakthroughs in key productive instruments have repeatedly expanded the frontier of what human societies could produce. The steam engine mechanized physical work and enabled the large-scale substitution of human and animal power [19]. Electrification and assembly-line systems reorganized production around standardized, continuous, and high-throughput manufacturing [20]. Later, computers and digital networks transformed how information was processed, calculations were performed, and communication and

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

coordination were organized, extending technological augmentation from manual labor to cognitive, administrative, and organizational work [21]. In each case, the importance of the new technology did not lie only in its technical novelty. More importantly, it changed how labor, capital, and coordination were organized at scale.

Despite their significance, a common feature of earlier productive instruments is that they mainly amplified specific aspects of human work rather than supporting broad execution across tasks. Industrial machines primarily strengthened and accelerated physical execution, while information systems and conventional software improved the speed, accuracy, and scale of calculation, storage, and rule-based processing [22,23]. These technologies generated substantial productivity gains, but their contribution was usually constrained by relatively fixed operating logics. Most were highly effective within predefined scopes and specific domains, while still depending on humans to set goals, coordinate workflows, and handle unexpected cases.

This distinction is central to understanding the potential economic role of agentic AI. Historically, the economic significance of novel technical systems lies not only in improving task performance, but also in reshaping the structure of work [18,24]. In the next subsection, we apply this lens to agentic AI and examine how recent advances in capability, cost, benchmarking, and deployment have made it a potential driver of productivity.

2.2. Agentic AI: From Capability to Productivity

Within the framework mentioned above, agentic AI can be understood as a new productive instrument. Unlike earlier digital tools that mainly supported information retrieval, rule-based processing, or isolated task assistance, agentic AI can increasingly participate in longer and more complete workflows. Built on large language models and related foundation-model capabilities, these systems can interpret task contexts, reason about goals, decompose problems into executable steps, invoke external tools, and interact iteratively with dynamic environments to complete non-trivial objectives [4–9,12,25]. Emerging systems such as `OpenClaw` [10] and `Claude Code` [11] already illustrate their growing role in practical workflows and developer ecosystems.

Compared with tools of previous eras, agentic AI exhibits two fundamentally transformative properties: **generality** and **autonomy**. Generality refers to the ability to operate across diverse tasks, domains, and knowledge contexts. While traditional software is highly effective within predefined procedures, agentic AI can be repurposed across linguistic, analytical, and operational workflows, making it a reusable asset across the production pipeline. Currently, autonomy enables these systems to complete closed-loop workflows with minimal human intervention. Instead of merely executing isolated commands, agentic systems maintain context, decompose goals, invoke tools, and adapt to dynamic feedback. This shifts their role from assisting with fragmented micro-tasks to taking ownership of complete units of work. Together, these properties make agentic AI a dual engine of productivity: an **efficiency mechanism** that compresses complex workflows, and an **expansion mechanism** that enables previously unfeasible forms of human-machine coordination.

Two primary technological drivers enable this shift from capability to productivity. First, the capabilities of agents have improved substantially. Modern agents can perform long-horizon planning with feedback-driven revision, conduct multi-step reasoning, maintain structured memory across interactions, decompose complex goals into executable sub-tasks, and reliably invoke external tools within dynamic environments [4,5,8,26,27]. Second, the cost of unit inference has decreased significantly [28]. Improved capabilities make it possible for agents to carry out more complex tasks in a stable manner, while lower inference costs make repeated and large-scale deployment economically feasible. Together, these changes allow agentic AI to move beyond experimental demonstrations and become part of real production workflows.

Recent macro-level signals also point to rapid expansion in investment and market expectations. The Stanford AI Index 2025 reports that U.S. private AI investment reached $109.1B in 2024, while global private investment in generative AI totaled $33.9B, representing an 18.7% year-over-year increase. Looking ahead, International Data Corporation (IDC) projects that global AI spending will

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

exceed $632B by 2028 [29], and Bloomberg Intelligence estimates that the generative AI market could approach $1.3T by 2032 under rapid growth assumptions [30]. Although investment is not itself productivity, these signals indicate that firms increasingly expect AI systems to become economically consequential production technologies rather than marginal digital tools.

In parallel with these investment trends, the evaluation of AI systems is also shifting from narrow capability benchmarks toward productivity-oriented assessment. New benchmarks such as GDPval replace automatic matching scores with blind pairwise comparisons by domain experts, who evaluate AI and human outputs on multi-hour, multi-document tasks that resemble real office work [31]. Long-horizon agent environments test whether agents can sustain performance over extended plan-and-execute scenarios. For example, Vending-Bench evaluates agents running a small business over many simulated days [32], while EcoGym tests agents in interactive economies such as freelance work and business operations over long horizons [33]. Other efforts directly examine the automation potential of high-value professional tasks. For instance, the AI Productivity Index (APEX) measures performance in consulting, finance, law, and medicine [34]. At the same time, the Remote Labor Index and RE-Bench focus on remote knowledge work and research tasks [35,36]. Taken together, these evaluations show a clear shift from measuring whether AI can perform narrow tasks to measuring whether it can reliably deliver economically valuable work at or above professional quality.

Early enterprise-level evidence also suggests measurable performance gains. Based on time estimates from approximately 100,000 AI-assisted conversational tasks, Anthropic reports that, for some tasks that take around 90 minutes without assistance, AI support can reduce completion time by approximately 80%. In knowledge-intensive services, a joint experiment by BCG and Harvard Business School finds that consultants using AI completed tasks 25.1% faster while achieving a 40% improvement in quality scores [37]. At the macro level, extrapolations under widespread adoption scenarios project that U.S. labor productivity growth could increase by 1.8 percentage points annually over the coming decade, with corresponding increases in Total Factor Productivity. Long-run GDP gains are estimated to range from 1.5% by 2035 to 3.7% by 2075 [38].

At the same time, the translation from technical capability to sustained productivity is not automatic. The economic impact of agentic AI depends not only on model performance and inference cost, but also on how these systems are embedded in organizational structures, decision processes, and existing workflows [39**?**]. In many settings, adoption remains gradual, and human oversight still plays a central role [21,40]. Understanding how micro-level performance improvements become broader economic value therefore requires more than isolated case evidence. It requires a structured account of how responsibility shifts from humans to AI systems, how this transfer changes workflow design, and how these patterns differ across industries.

The key analytical problem is therefore a conversion problem: when and how do model capabilities become durable productivity gains in real workflows? This question links the conceptual lens above to the industry analysis that follows, where the same technical capability can produce very different productivity outcomes depending on workflow embedding, verifiability, error tolerance, and institutional accountability.

In the following sections, we formalize these productivity gains and introduce a four-stage framework to describe the transfer of agency from humans to AI systems. We then use this framework to examine the current pattern of industrial diffusion and explain why the level of AI autonomy differs across domains.

3. Framework: Mechanisms and Delegation Stages

3.1. Definition of AI for Productivity

In this paper, **AI for Productivity** refers to the use of agentic AI in real workflows to either reduce the human effort required for existing output or make valuable work feasible that would otherwise remain too costly, slow, or complex. Like earlier productivity technologies, Agentic AI can raise productivity through two mechanisms across industries. It can reduce the time and labor

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

required for existing work. It can also make economically valuable work feasible that would otherwise remain undone because it is too costly or too complex, thereby further freeing human capacity, as we discuss next

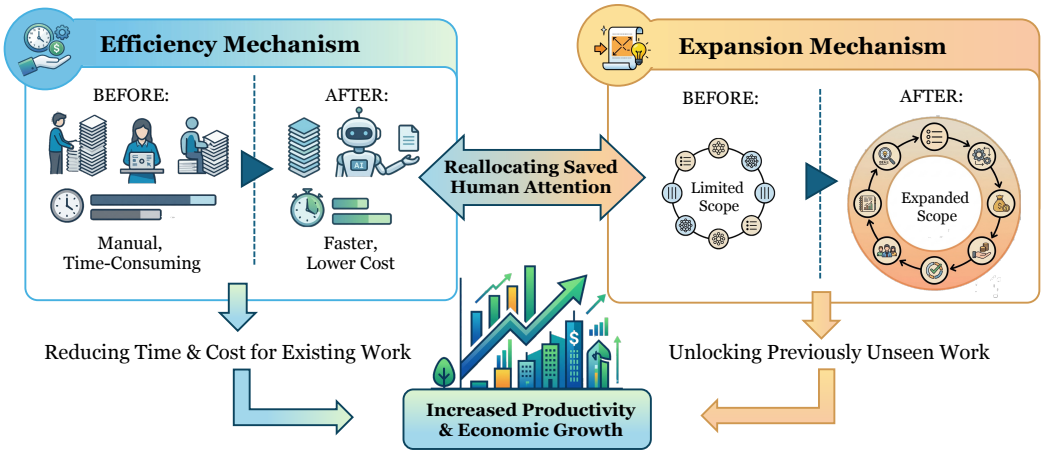


Figure 2. An illustration of the two mechanisms. Efficiency Mechanism: It enhances productivity by reducing time and cost for existing work. Expansion Mechanism: It expands the task scope and unlocks the unseen work to foster productivity.

The Efficiency Mechanism

The Efficiency Mechanism is based on reducing the time and labor required for existing work. It applies when the cost of running an agent system for a task becomes lower than the cost of human time for the same task. In that case, organizations can automate parts of existing workflows and complete the same work faster and at lower cost [41]. This leads to immediate and measurable improvements in execution speed through task acceleration and the automation of repetitive work. Empirical evidence supports sizable time savings from Agentic AI assistance in knowledge work [37,42]. Research from Anthropic indicates that Agentic AI can accelerate individual tasks by up to 80%, particularly in knowledge-intensive sectors. At the macro level, Wharton's Penn Wharton Budget Model (PWBM) projects that this efficiency boost will increase aggregate GDP and productivity levels by 3.7% by 2075 [38], generating a transient but powerful wave of Total Factor Productivity (TFP) growth. Under this mechanism, productivity is gained by doing the same economic tasks, but with substantially lower human expenditure.

The Expansion Mechanism

While efficiency is about reducing the time and labor required for existing work, the Expansion Mechanism is about unlocking work that would otherwise remain undone. It relies on the autonomy and scalability of agentic AI to make previously costly or complex tasks feasible. Recent studies suggest that a significant share of AI-assisted work would not have been done at all without AI assistance [43,44]. For example, in Anthropic's internal study, about 27% of AI-assisted tasks were described by users as work they "would not have done at all" without LLMs. Such work often includes exploratory iteration, large-scale engineering completion, and many long-tail improvements that are valuable but usually left undone because teams lack time or attention [42]. In practice, reductions in time and labor costs can also make this broader scope of work easier to sustain and scale. Therefore, "AI for Productivity" should not be understood only as automation or substitution. Its more important long-run effect may come from reallocating saved human attention and labor toward higher-value innovation, decision-making, and service, which can drive continuous improvement in both products and workflows [45].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

3.2. Delegation Stages from a Productivity Perspective

To systematically examine how AI reshapes productivity across industries, we classify agentic AI deployment by the depth of production responsibility delegated from humans to AI systems. These stages do not simply describe how AI systems evolve; they describe how execution, coordination, monitoring, and eventually strategic control are redistributed within production workflows.

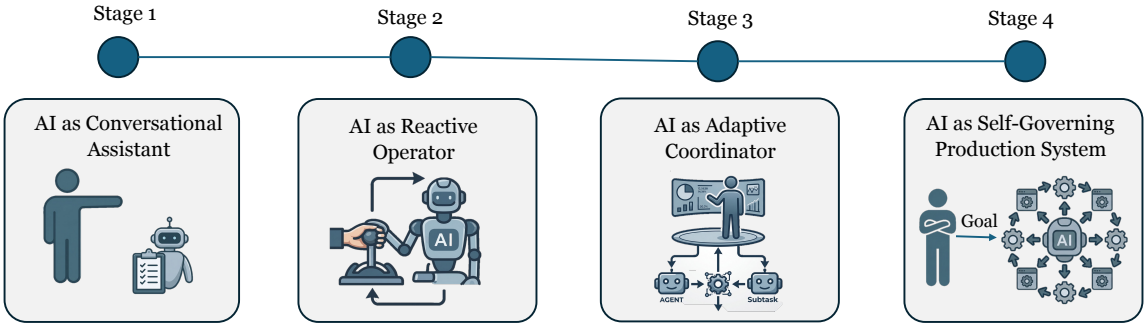


Figure 3. Four stages of AI reshaping productivity.

Stage 1: AI as Conversational Assistant

Stage 1 marks the baseline of AI integration, where AI systems function as responsive, informational instruments. At this stage, AI leverages its generality to assist with information retrieval, drafting, and preliminary analysis under single-turn instructions. The productivity gains are primarily driven by the *Efficiency Mechanism*, where AI accelerates the early phases of work by providing initial ideas.

However, as a conversational assistant, AI cannot interact with the external environment, maintain persistent state, or participate in the actual execution of tasks. Furthermore, its preliminary suggestions are often incomplete or unreliable. The productivity ceiling of this stage is thus defined by the absolute reliance on manual execution and output validation. Humans must fully take over to verify the accuracy of the AI's advice, translate these rough ideas into concrete actions, and manually drive the entire production loop from start to finish.

Stage 2: AI as Reactive Operator

Stage 2 advances to operational automation, where AI transitions from a passive assistant to a reactive digital operator. At this stage, systems can autonomously execute narrowly scoped operations such as structured data manipulation, code generation, or text refinement. The *Efficiency Mechanism* deepens as productivity gains shift from preparatory assistance to the direct substitution of repetitive, low-level execution. By automating this operational labor, AI deeply integrates into the individual steps of the workflow.

Yet, systems in Stage 2 remain severely constrained by their limited contextual scope and inability to navigate dynamic environments. They cannot decompose complex goals or seek alternative solutions when confronted with non-standard inputs and errors. The productivity ceiling of this stage is defined by the constant need for human oversight and error correction. While humans no longer perform the discrete tasks themselves, they must constantly intervene to provide highly specific instructions, check intermediate outputs, stitch together isolated AI outputs, and fix operational mistakes to prevent the workflow from breaking.

Stage 3: AI as Adaptive Coordinator

In Stage 3, AI acts as an adaptive coordinator. Beyond simple execution, these systems decompose complex goals into workflows, utilizing persistent memory and dynamic tool-integration to manage extended tasks and recover from failures. At this stage, the *Efficiency Mechanism* deepens as AI automates procedural coordination, reducing marginal costs and scaling individual capability to a

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

team level. By removing high coordination overhead, AI unlocks complex economic activities that were previously unfeasible or uneconomical. This shift in efficiency thus enables the *Expansion Mechanism* take effect.

Despite these advancements, AI in Stage 3 remains bounded by explicitly provided contexts and predefined success criteria. While it can efficiently execute a well-structured workflow, it cannot understand implicit intentions or autonomously structure ambiguous concepts into solvable projects. The productivity ceiling of this stage is therefore defined by the inherent burden of problem scoping and task specification. Humans must still manually translate open-ended goals into detailed execution blueprints and define precise requirements before the AI can take over the coordination.

Stage 4: AI as Self-Governing Production System

In Stage 4, AI evolves into an autonomous self-governing production system. Moving beyond task execution, these systems translate abstract, outcome-oriented goals into structured subgoals through multi-dimensional trade-offs. They operate in open environments, maintain consistent internal state, adapt to dynamic constraints, and orchestrate multiple agents as a unified production ecosystem. Building on further improvements in the *Efficiency Mechanism*, the *Expansion Mechanism* reaches its full potential. By enabling scalable, asynchronous decision-making beyond human attention limits, AI autonomously organizes activities that were previously too complex or costly to manage, effectively expanding the production frontier.

However, giving AI this level of autonomy introduces new bottlenecks related to goal alignment and data limitations. Since the system relies solely on quantifiable information, it remains blind to complex human factors and unpredictable real-world events. The productivity ceiling of this stage is defined by the systemic overhead of strategic governance and alignment verification. While eliminating human execution and coordination provides massive productivity gains, it creates a new requirement for humans to monitor AI decisions, build safety measures, and ensure that high-level strategies remain aligned with human goals.

Taken together, these four stages trace a progressive delegation of production responsibility from humans to AI systems. Across this trajectory, the *Efficiency Mechanism* continuously deepens to optimize existing economic activities, which ultimately allows the *Expansion Mechanism* to take effect, enabling AI to actively expand the boundaries of industrial capability.

4. Evidence: From Deployed Systems to Productivity Measurement

4.1. Deployed Systems as Evidence of Productivity Conversion

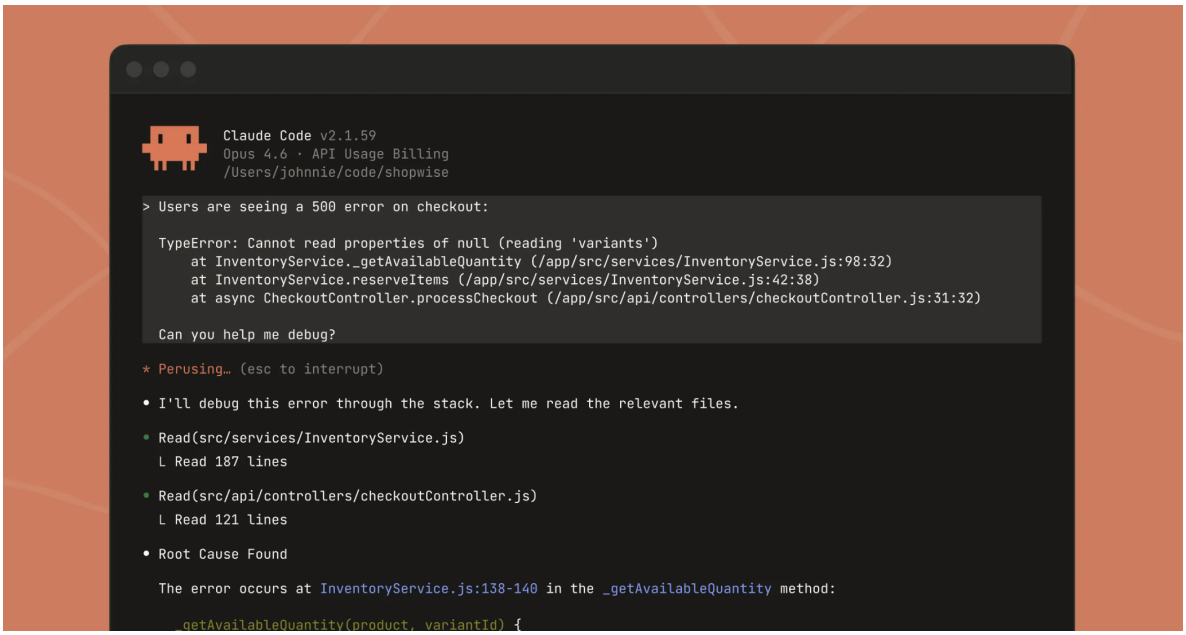
To illustrate how theoretical stages of agentic AI are converted into real-world productivity gains, we examine three pioneering AI productivity tools: Claude Code [11], Claude Cowork [46], and OpenClaw [10]. These cases are not merely examples of product functionality. They show how workflow context determines whether agentic capability becomes sustained productivity: executable feedback in software engineering, data-workflow access for non-technical workers, and persistent background operation for individual users.

Claude Code

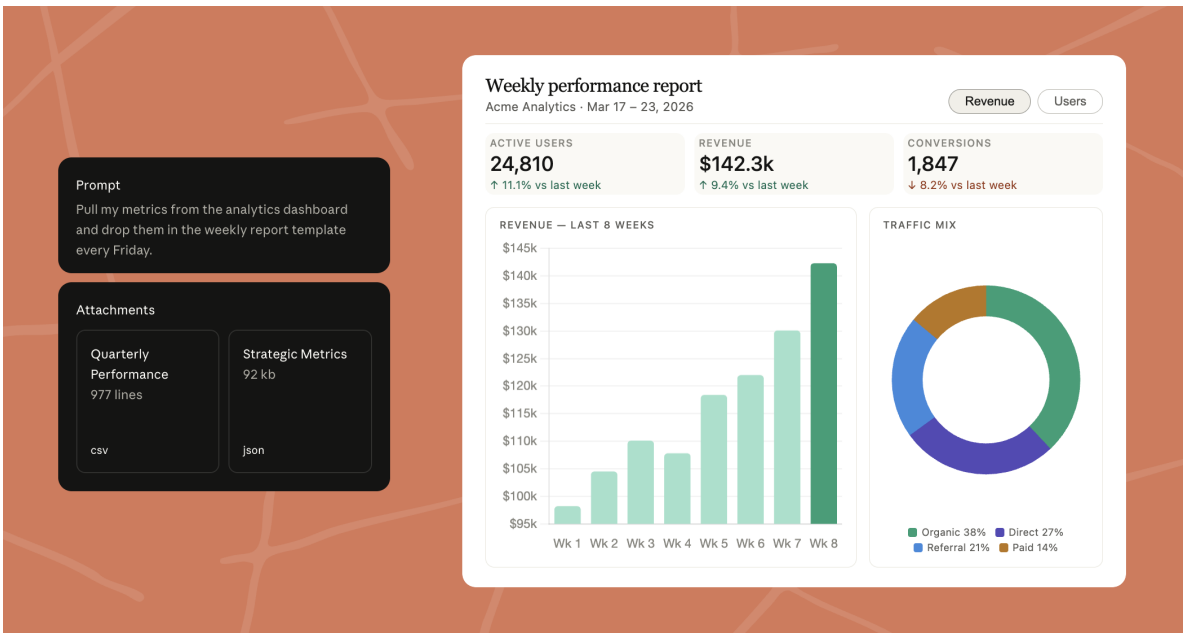
Claude Code shows why software is an early frontier for agentic productivity: the work is digital, feedback is executable, and iteration loops can be closed inside the development environment. As a highly capable autonomous programming agent designed for complex software engineering, it is integrated deeply into native development environments and functions not only as an assistant, but as a powerful, composable command-line engine. It uses persistent memory to actively master project-specific architectures and user conventions, with reusable skills enabling it to leverage prior experience for continuous self-evolution. It also supports multi-agent collaboration for robust, parallel task execution across large-scale systems and balances autonomy with strict permission controls to prevent unauthorized local file modifications.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1



(**a**) Claude Code [47]



(**b**) Claude Cowork [46]

Figure 4. Evidence from AI Productivity Tools. Left: Claude Code autonomously debugging a local repository by tracing error stacks and navigating the source code within a terminal environment. Right: Claude Cowork synthesizing fragmented data from local CSV and JSON files into a performance report with visualizations, executed as a scheduled task to ensure continuous, automated reporting on a predefined cadence

These capabilities enable Claude Code to operate through both efficiency and expansion mechanisms. Under the *Efficiency Mechanism*, it drastically accelerates the software development lifecycle. For example, Rakuten reported that Claude Code performed 7 hours of continuous autonomous coding with 99.9% accuracy on complex refactoring tasks, reducing the time required to release new features by 79% [48].

Furthermore, it triggers the *Expansion Mechanism* for both engineers and non-engineers. For engineers, it activates large-scale projects that might otherwise be delayed or abandoned due to resource constraints. At Wiz, a 50,000-line Python library was migrated to Go in 20 hours, a task originally estimated to take several months. According to the company's leadership, this project would

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

not have been pursued without Claude Code [49]. Additionally, the system lowers the barrier to software creation for non-technical roles, such as founders and product managers, by allowing them to develop functional prototypes from natural language descriptions.

Consequently, Claude Code not only accelerates existing software development processes, but expands the boundaries of who can create software and what projects can realistically be undertaken.

Claude Cowork

Claude Cowork illustrates how agentic AI expands productivity by giving non-technical workers access to data workflows that previously required specialized technical labor. Claude Cowork [46] applies the agentic architecture of Claude Code to non-technical knowledge workflows. Designed as a desktop-based assistant for enterprise productivity, it can produce professional deliverables such as spreadsheets and slides that can be further edited with Excel and PowerPoint. Users can also automate recurring duties via scheduled tasks and utilize project workspaces to provide isolated environments for distinct tasks.

These capabilities allow Cowork to automate high-effort and repeatable work, such as organizing local files, preparing meeting materials, and managing schedules, thereby improving productivity through the *Efficiency Mechanism*. At Airtree, Cowork reduced market and competitor research from two days to minutes and cut board meeting summaries from 30 minutes to two, allowing the team to spend more time analyzing insights and refining their strategy rather than creating reports [50].

Furthermore, it triggers the *Expansion Mechanism* by empowering non-technical professionals to build systems that support their decision-making process. At Zapier, generating a unified visualization of influencer marketing data was previously unfeasible, as critical metrics like investment data and real-time engagement were scattered across multiple technical systems. The required SQL complexities and cross-platform integrations fell beyond the marketing team's technical scope, forcing them to rely on delayed, retrospective reporting. With Cowork, a non-technical marketing leader independently built a live dashboard that synthesizes these fragmented data streams into a daily-updated visualization. This transition from retrospective analysis to forward-looking guidance allows the team to proactively reallocate budgets and prioritize high-performing channels [51].

In this way, Cowork not only accelerates existing workflows but also creates new analytical practices and decision-making processes that were once precluded by technical barriers.

OpenClaw

OpenClaw highlights a shift from one-time assistance to persistent background operation, where productivity gains come from continuous monitoring and autonomous task initiation. Beyond professional scenarios, frameworks like OpenClaw [10] exemplify a shift toward deeply personalized autonomous agents. Deployed directly on local machines, OpenClaw securely accesses private documents and interacts with native applications to establish a highly specific personal context. The system incorporates a persistent memory architecture, allowing the agent to adapt to user habits and historical data over time. Instead of reacting passively, OpenClaw monitors system states periodically via a heartbeat mechanism, initiates tasks autonomously, and supports recurring operations through cron-based scheduling. Users can monitor and direct this persistent local engine through integration with standard messaging platforms, including WhatsApp, Telegram, and Discord.

These capabilities enhance productivity through the *Efficiency Mechanism* by reducing the need for active human monitoring. For example, by leveraging the heartbeat mechanism, OpenClaw can continuously monitor deployed local services or long-running machine learning training tasks. When an anomaly or crash occurs, OpenClaw autonomously diagnoses the issue and executes necessary repairs, entirely without human intervention. The user simply receives a brief recovery notification via Telegram, transforming tedious troubleshooting into seamless automation.

At the same time, OpenClaw drives the *Expansion Mechanism* by allowing solo professionals to deploy continuous, enterprise-grade operational systems from their personal computers. For example, a solo entrepreneur can establish a 24/7 market intelligence pipeline that autonomously scrapes

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

industry databases overnight, filters massive datasets based on the user's highly specific business context, and pushes actionable, synthesized leads directly into a WhatsApp chat by morning. By transforming a static local machine into an autonomous digital back-office, OpenClaw empowers individuals to execute data-intensive, continuous ventures that would previously require hiring a dedicated team of analysts.

Together, these case studies illustrate how agentic AI translates autonomy into real productivity gains. Across domains, these tools show that agentic AI not only accelerates existing workflows but also expands the scope of what individuals and organizations can achieve.

4.2. Measuring Productivity: From Capability Scores to Economic Work

The emergence of productivity-oriented benchmarks indicates a broader measurement shift. AI systems are increasingly evaluated not only by whether they answer correctly, but by whether they can complete economically valuable work over realistic time horizons. They tend to take one of two stances. The first targets the economic value an agent can produce. The second measures the length of task an agent can autonomously complete. Several benchmarks adopt the first stance. OpenAI's GDPval scores models head-to-head against industry experts on 1,320 tasks drawn from 44 occupations across the nine largest sectors of U.S. GDP [31]. Mercor's APEX uses rubric-scored expert tasks in investment banking, consulting, law, and medicine, with an agentic variant (APEX-Agents) for multi-step execution [34]. $OneMillion-Bench prices 400 expert-curated tasks at real market wages across five professional domains [52]. Xbench frames productivity as Technology-Market Fit, weighting performance by cost and market demand in verticals like recruitment and marketing [53]. Vending-Bench asks agents to run a simulated business for a year and scores them solely on dollars earned [32]. The second stance is best represented by METR's time horizon [54], which collapses productivity to a single scalar: the length of human-expert work an agent can complete with 50% success. RE-Bench [36] takes a similar approach for ML research engineering tasks, evaluating agents against 8-hour human-expert attempts.

Constructing such evaluations is resource-intensive. Each requires domain-expert knowledge. GDPval's tasks come from professionals averaging 14 years of experience. APEX hires former Goldman analysts and big-law associates. $OneMillion-Bench took over 2,000 hours of expert collaboration. Each also requires long task execution. A single Vending-Bench run consumes 60 to 100 million tokens, and METR tasks range up to 16 hours of human time. Indeed, for most industries, the role of agentic AI in driving productivity remains without formal evaluation. Benchmarks capture only part of productivity because they are better at measuring task performance, economic output proxies, and time horizons than workflow embedding, responsibility delegation, and institutional accountability. The industry analysis therefore complements benchmark evidence by assessing where agentic AI is actually embedded in workflows and how much production responsibility has been delegated.

4.3. Research Scope to Industry Diffusion

Our goal is not to exhaustively catalog every AI application in each sector. Instead, we use representative academic studies, deployed systems, benchmarks, and industry evidence to assess the depth of agentic AI integration.

We adopt the same industry list identified in the OpenAI GDPval benchmark [31], which focuses on sectors with the largest contributions to U.S. GDP and high-earning, predominantly digital occupations. We then categorize these industries into three groups by assessing whether AI is embedded in real workflows, whether it can execute multi-step tasks, whether it uses external tools and feedback loops, whether human responsibility is partially transferred to AI systems, and whether observed productivity gains primarily reflect efficiency or expansion.

Specifically, in Section 5, we discuss **nascent-diffusion** industries, including *manufacturing, agriculture, real estate, and government*, which remain largely concentrated in Stage 1 and early Stage 2. In Section 6, we examine **growing-diffusion** industries such as *trade, media, education, and legal services*, which increasingly demonstrate Stage 2 and partial Stage 3 capabilities. In Section 7, we analyze

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

widespread-diffusion industries including *information technology, financial services, and medicine*, which show the most active exploration of Stage 3 autonomy and early Stage 4 traits in constrained settings.

5. Nascent-Diffusion Industries

5.1. Manufacturing: Bounded Automation in Safety-Critical Production

Manufacturing represents a high-risk physical system characterized by strong production continuity, high downtime costs, and strict safety constraints. These conditions make it more feasible for Agentic AI to initially augment consultative and decision-support tasks rather than achieve end-to-end automation. Over the past decade, industrial AI has followed a clear trajectory: from localized perception and defect detection, to knowledge-enhanced systems via Retrieval-Augmented Generation (RAG), and further toward integration with engineering toolchains and early multi-agent architectures.

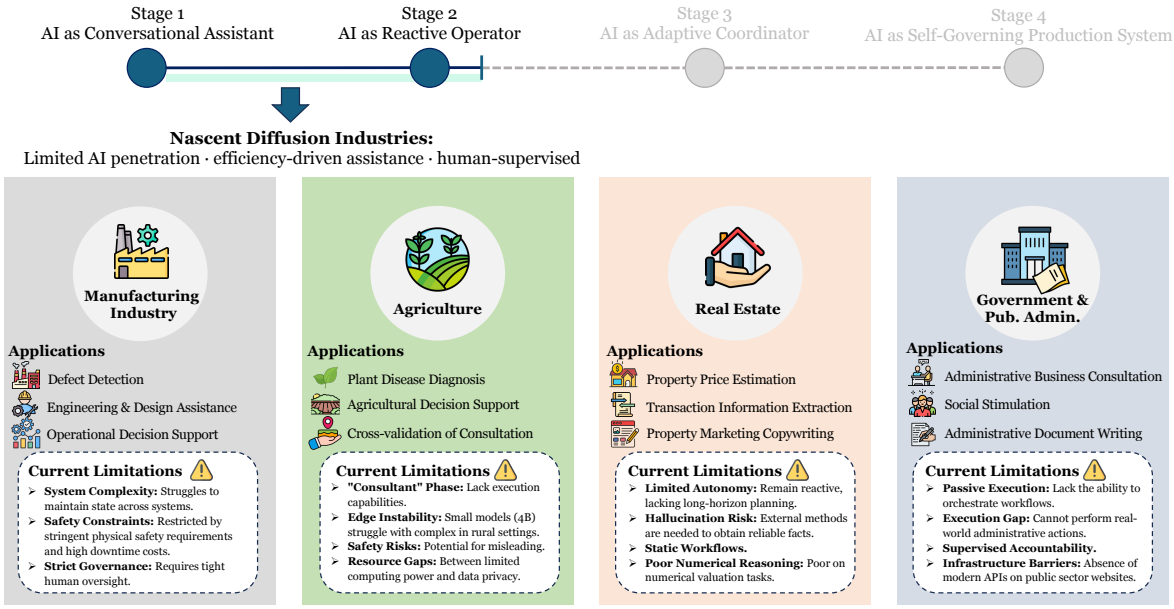


Figure 5. Nascent Diffusion Industries at Stages 1–2 of AI Adoption. These sectors share traits of limited AI penetration, efficiency-driven assistance, and reliance on human supervision. For each, the figure identifies representative applications alongside key limitations, such as system complexity, hallucination risk, execution gaps, and infrastructure barriers, that hinder progression toward more autonomous Stage 3 and 4 deployments.

Diagnosis and Engineering Support as the Entry Point

Early industrial AI was dominated by single-task discriminative models such as anomaly detection and predictive maintenance, constrained by limited abnormal samples and high false-positive costs. Representative methods include PaDiM [55] and PatchCore [56] on the MVTec AD benchmark [57]. While these approaches achieved perception-level automation, they lacked engineering context and generalization across production systems. With the accumulation of enterprise knowledge (e.g., manuals, logs, and specifications), industrial AI evolved toward RAG-based paradigms that connect models with private data sources, improving traceability and reducing hallucinations. Systems such as Schneider Electric's EcoStruxure Copilot [58] exemplify this trend, enabling natural-language fault analysis and maintenance recommendations. However, these systems remained primarily limited to operational support.

Controlled Agents for Partial Operational Workflows

As automation complexity increased, generative AI expanded upstream into engineering design. "Engineering copilots" integrated with industrial platforms began supporting code generation, documentation retrieval, and control logic explanation. For example, Siemens Industrial Copilot [59] enables natural language–driven automation code generation and system configuration. Academic

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

efforts further explore generating PLC programs and simulation models from natural language via retrieval-augmented and preference-aligned training. Despite these advances, such systems largely function as cognitive assistants, lacking continuous workflow awareness and the ability to execute cross-system tasks autonomously.

Recent developments extend AI from engineering to operations by integrating real-time industrial data. Systems such as ABB Genix Copilot [60] combine sensor data and historical records to interpret alarms and support decision-making. More advanced architectures, explored by companies like Honeywell [61] and Siemens [62], introduce multi-agent coordination and tool invocation for partial task automation. However, due to strict requirements on safety, auditability, and control, these systems operate under human governance and remain at the stage of "interpretation–recommendation–partial automation," lacking full autonomy.

Overall, from 2016 to 2025, manufacturing AI has evolved from perception-level automation to early-stage Agentic systems, progressively expanding its functional scope from defect detection and predictive maintenance to knowledge-grounded reasoning, engineering assistance, and partial workflow orchestration. Modern systems are increasingly capable of integrating multi-source industrial data, retrieving enterprise knowledge via RAG, assisting in code generation and system configuration, and providing interpretable recommendations for operational decision-making. Market forecasts further reflect this trajectory, with MarketsandMarkets projecting growth from USD 34.18B in 2025 to USD 155.04B by 2030 (CAGR 35.3%) [63], and Technavio estimating a 37.7% CAGR for generative AI in manufacturing [64].

Why Production Control Remains Hard to Delegate

However, despite these advances, the autonomy boundary of industrial AI remains fundamentally constrained by physical safety requirements, production continuity, and the high cost of errors. Current systems still lack persistent state awareness, robust cross-system execution capability, and deterministic reliability in long-horizon tasks. As a result, most applications remain confined to roles such as interpretation, diagnosis, and recommendation, with only limited forms of tool-mediated automation. From the perspective of autonomy, manufacturing AI is still predominantly situated in AI-as-Conversational-Assistant and AI-as-Reactive-Operator regimes, where AI enhances human decision-making rather than independently controlling end-to-end production processes.

5.2. Agriculture: Decision Support under Environmental Uncertainty

Agriculture is a highly variable physical production domain shaped by environmental uncertainty, biological heterogeneity, and dispersed field conditions. These characteristics make it difficult for Agentic AI to move directly toward end-to-end autonomous execution in real farming scenarios. As a result, LLMs are increasingly applied in agriculture first for tasks such as yield prediction, disease diagnosis, and farm management, where they can support decision-making and selected operational steps. To overcome the domain knowledge limitations of general-purpose models, agricultural AI has evolved along a clear technical trajectory: from specialized unimodal LLMs to multimodal integrations and currently toward agentic systems augmented by tools. The following analysis assesses these advances as a transition from passive consultation to more capable but still bounded decision-support systems

Agronomic Knowledge Support as the Entry Point

Specialized agricultural LLMs are transitioning from parametric adaptation toward dynamic retrieval and data synthesis. While early models like AgriBERT [65] and AgroGPT [66] leveraged domain-specific fine-tuning to improve terminology alignment, subsequent systems such as AgroLLM [67] and ShizishanGPT [68] introduced RAG to provide reliable, localized pesticide recommendations and precise plant disease diagnoses by dynamically retrieving the latest agricultural guidelines. More recently, IPM-AgriGPT [69] and AgriGPT [70] have explored data-centric strategies, utilizing adversarial generation and multi-agent pipelines to synthesize high-quality agronomic corpora, addressing the critical scarcity of expert agricultural data. These advancements suggest that coupling data

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

synthesis with dynamic retrieval is an efficient pathway for adapting general-purpose LLMs to the agricultural domain.

Multimodal and Tool-Augmented Agents in Field Workflows

Substantial research efforts have been directed toward extending large models to multimodal inputs, a critical requirement for visual and hands-free agricultural operations. Early vision-language initiatives, such as Agri-LLaVA [71], focused on image-based crop disease diagnosis, while subsequent efforts like the AgriGPT-VL Suite [72] further refined multimodal alignment via curriculum learning and policy optimization. To address the demands of real-world field environments, research has expanded into tri-modal architectures. For instance, AgriGPT-Omni [73] integrates speech, text, and vision to handle complex field acoustics and dialects. These collective advancements represent a pivotal leap toward integrated multimodal interaction frameworks, significantly unlocking the latent productivity of digital agriculture.

To bridge the gap between passive information retrieval and actionable decision support, recent research efforts have integrated LLMs into specialized tool-use and multi-agent frameworks. Systems such as AgriDoctor [74] exemplify this trend by dynamically orchestrating vision models and knowledge graphs for precision management. Further research efforts, such as the AgMMU framework [75], utilize collaborative agent pipelines—comprising specialized Retriever, Reflector, and Answerer agents—to maintain reasoning consistency across diverse visual inputs. Moreover, this research trajectory is currently culminating in physical execution. Specifically, hierarchical planners such as AgriAgent [76] and sensor-integrated systems like CottonBot [77] enable the automation of robotic workflows and irrigation, thereby signaling a definitive paradigm shift toward autonomous precision agriculture.

Why Farm-Level Autonomy Remains Fragile

Overall, agricultural AI has evolved from specialized LLMs to retrieval-enhanced systems, multimodal models, and early agentic architectures with tool-use capabilities. These systems increasingly combine agronomic knowledge, multimodal perception, and selected forms of planning or robotic support to improve diagnosis, recommendation, and partial workflow execution. However, agricultural generative AI still remains largely in *Stage 1 (Conversational Assistant)*, functioning primarily as advisory and diagnostic systems that require human intervention to translate model outputs into physical actions. Recent agentic frameworks suggest an emerging transition toward *Stage 2 (Reactive Operator)*, but current models still lack the robustness, long-horizon planning, and reliable decision-making needed for dynamic agricultural environments. Key bottlenecks include the reasoning limitations of small language models deployed at the edge and the lack of tightly integrated software and hardware infrastructures required for complex physical operations. As a result, most agricultural applications remain assistant-like or narrowly bounded operator systems rather than autonomous managers of end-to-end farm operations.

5.3. Real Estate: Document Intelligence without Transaction Autonomy

Real estate is a document-intensive and coordination-heavy domain that relies on fragmented information, localized market knowledge, client communication, and multi-party transactions. These characteristics make it more feasible for Large Language Models to first augment information processing and decision-support tasks rather than move directly toward end-to-end autonomous execution. In this context, research efforts have focused on integrating LLMs into traditional workflows to handle unstructured data and improve selected stages of real estate analysis, contract processing, and marketing. From a technical perspective, the evidence broadly reflects a progression from in-context learning, to supervised finetuning, and further toward Retrieval-Augmented Generation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Documents and Market Information as the Entry Point

The initial application of LLMs in the real estate domain primarily leverages their intrinsic semantic understanding capabilities. As an early exploration, researchers have applied LLMs with in-context learning for housing price prediction. Although empirical evaluations demonstrate that isolated LLMs underperform traditional machine learning models on price prediction tasks [78], these efforts establish a baseline for domain-specific deployment. Similar approaches have also been utilized to evaluate the attribute extraction performance of LLMs across various prompting configurations (eg. one-shot, few-shot, etc.) [79]. At this stage, however, LLMs function mainly as flexible advisors for narrow tasks.

Grounded Generation for Client-Facing Workflows

To handle domain-specific data processing needs, research has progressed toward utilizing fine-tuned LLMs for automated information extraction. These systems are designed to read unstructured real estate contracts and answer certain questions accordingly [80]. Compared to rule-based approaches, instruction-tuned models demonstrate superior robustness and reduce the need for manual feature engineering. This development expands the role of LLMs from general semantic assistance to more structured document understanding, but their function remains largely supportive, centered on contract reading and information processing rather than autonomous transaction execution.

Recent applications of Retrieval-Augmented Generation in real estate leverage retrieved data to enhance the persuasiveness of marketing content. For instance, the AI Realtor framework [81] utilizes RAG to extract unexpected features associated with a property. This targeted grounding strategy effectively captures user interest and stimulates purchase desire, demonstrating RAG's capability to generate highly engaging and persuasive marketing content. Even so, these systems remain concentrated in search, description, and marketing support, rather than in higher-stakes functions such as negotiation, compliance management, or multi-party coordination.

Why Transaction Responsibility Remains Human-Centered

Overall, real estate AI has evolved from early prompt-based semantic assistance to fine-tuned models for contract understanding and retrieval-grounded systems for marketing and client interaction. However, despite this technical progression from zero-shot inference to RAG-based systems, most applications still remain within *Stage 1 (AI-as-Conversational-Assistant)* or early *Stage 2 (AI-as-Reactive-Operator)*. Current systems can support tasks such as attribute extraction, document interpretation, price-related analysis, and grounded property description, but they remain largely reactive and lack persistent state, long-horizon planning, and autonomous task decomposition capabilities. In practice, the LLM functions mainly as an advanced automation component embedded within a fixed workflow, while core decisions such as pricing strategies, negotiation approaches, legal judgment, and transaction execution remain the responsibility of human actors. As a result, real estate AI still operates primarily in assistant-like roles focused on information processing, document understanding, and client-facing support, rather than as an autonomous manager of end-to-end real estate transactions.

5.4. Government and Public Administration: Public-Service Support

Government and public administration constitute a high-stakes institutional domain shaped by legal accountability, procedural rigidity, public trust, and heterogeneous legacy systems. These characteristics make it difficult for large language model-based agents to move directly toward autonomous decision-making or delegated public authority. As a result, current research has more often introduced such systems first in roles related to regulatory retrieval, scenario simulation, and bounded administrative assistance. From a technical perspective, existing work broadly reflects a progression from retrieval-augmented systems for regulatory compliance, to multi-agent social simulation systems for policy modeling, and further toward operational agentic systems for selected administrative workflows.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Regulatory Retrieval and Administrative Drafting as the Entry Point

A prevalent line of research in this domain focuses on mitigating hallucinations and enhancing regulatory alignment through Retrieval-Augmented Generation (RAG) and knowledge graphs (KGs). For instance, systems such as GraphCompliance [82] and PrivComp [83] integrate unstructured regulatory text with structured logic, allowing the model to validate its outputs against specific legal clauses. Technically, these systems prioritize precision over autonomy, acting as advanced semantic search engines with limited generative agency.

Simulation and Browser Agents for Bounded Public Workflows

A second stream employs multi-agent systems to model social dynamics. Frameworks like GOV-SIM [84] utilize game-theoretic scenarios to simulate collective behavior and resource governance, offering insights for resource governance strategies under varying communication constraints. Similarly, Park et al. [85] demonstrate that agents grounded on human interviews can better reflect public opinion for policy evaluation. Despite these improvements, existing research questions whether such agents are capable of reliable real-world decision-making, as they exhibit severe internal inconsistencies and misalignment with established evaluation frameworks [86]. At this stage, therefore, multi-agent systems are more useful as exploratory or analytical components than as dependable substitutes for public judgment.

A third category encompasses agentic systems engineered to execute specific administrative workflows. For administrative document generation, Musumeci et al. [87] employ a multi-agent framework that decomposes complex drafting tasks into collaborative sub-tasks for the production of administrative documents. Meanwhile, preliminary studies such as Agents4Gov [88] adapt conventional browser agents to accommodate the low-resource characteristics and absence of API functionality inherent to government website usage. These efforts extend agentic AI from policy support and simulation toward selected forms of workflow execution. However, the systems remain bounded to narrow administrative settings and may still require human oversight to ensure procedural correctness and operational reliability.

Why Public Authority Remains Hard to Delegate

Overall, government-facing AI has evolved from retrieval-grounded regulatory assistance to multi-agent policy simulation and early agentic systems for selected administrative workflows. Yet despite this architectural diversity, most applications still remain within *Stage 1 (AI-as-Conversational-Assistant)* and early *Stage 2 (AI-as-Reactive-Operator)*. Retrieval-augmented systems function mainly as fixed compliance-support pipelines rather than autonomous agents, while simulation agents serve as analytical components. Operational systems, such as browser navigation and administrative document drafting, support only localized sub-task execution and remain unable to perform workflow orchestration, or delegated public administration. As a result, current government AI systems still function primarily as specialized tools under human supervision rather than as autonomous entities entrusted with public administrative processes.

5.5. Agentic AI as Bounded Assistance in Early-Diffusion Domains

Across manufacturing, agriculture, real estate, and government, Agentic AI has begun to create visible value, but its role in production remains primarily that of an *assistant* rather than an autonomous executor. Recent progress has introduced domain-specialized fine-tuned models, RAG-enhanced advisory assistants, and early tool-using or multi-agent systems into real workflows. Yet in most cases, these systems still operate by supporting human decision-making, providing recommendations, and participating only in a small number of narrowly bounded execution steps. Adoption is therefore already visible across a relatively wide range of tasks, but the *depth* of delegation remains shallow. These sectors are thus representative *nascent-diffusion industries*: Agentic AI is no longer confined to isolated predictive or assistive use, but its integration into production remains uneven, highly

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

supervised, and largely concentrated in ***Stage 1 (AI-as-Conversational-Assistant)*** and early ***Stage 2 (AI-as-Reactive-Operator)***.

At this stage, productivity gains arise overwhelmingly through the ***Efficiency Mechanism***. By reducing human time in interpretation, drafting, diagnosis, and routine coordination, Agentic AI improves existing workflows without fundamentally reshaping their underlying control structure. By contrast, the ***Expansion Mechanism*** remains largely muted, because many tasks beyond routine human execution are constrained less by cognitive cost than by physical conditions, natural uncertainty, execution risk, and the need for direct intervention in real environments. This also explains why these sectors have not yet moved to the next stage of diffusion. Compared with digital-native domains, they are more tightly coupled to the physical world and less deeply embedded in fully digital operating environments, while placing much stricter demands on safety, low-risk execution, and verifiable and governable action. Since current agent still face limitations in persistent state, long-horizon reliability, and robust interaction with complex real-world production settings, broader delegation remains difficult to justify in practice. As a result, Agentic AI in these industries continues to function mainly as a supervised productive instrument for bounded assistance rather than as a fully autonomous production actor.

Table 1. Representative domain-specific applications in Nascent-diffusion industries.

Paper	Task	Technique	Stage
Manufacturing			
PaDiM [55]	Defect detection	Anomaly detection	Stage 1
PatchCore [56]	Defect detection	Anomaly detection	Stage 1
EcoStruxure Copilot [58]	Fault diagnosis	RAG	Stage 1–2
Siemens Industrial Copilot [59]	Engineering assistance	Tool-augmented LLM	Stage 2
ABB Genix Copilot [60]	Decision support	Industrial copilot	Stage 2
Honeywell AI Assistant [61]	Task automation	MAS	Stage 2–3
Siemens AI Automation System [62]	Workflow automation	MAS	Stage 2–3
Agriculture			
AgroLLM [67]	Management Q&A	RAG	Stage 1
ShizishanGPT [68]	Agricultural assistance	Tool-augmented LLM	Stage 1–2
Agri-LLaVA [71]	Pest / disease diagnosis	Multimodal LLM	Stage 1
IPM-AgriGPT [69]	Pest / disease management	Contextual reasoning	Stage 1
CottonBot [77]	Irrigation advice	RAG + agentic LLM tools	Stage 2
AgriAgent [76]	Task planning	Tool orchestration	Stage 2–3
Real Estate			
Utilizing Large Language Models for Information Extraction from Real Estate Transactions [80]	Information extraction	SFT	Stage 1
Real Estate Attribute Value Extraction Using Large Language Models [79]	Information extraction	Prompt engineering	Stage 1
AI Realtor [81]	Automated copywriting	RAG	Stage 1
On the Performance of Large Language Models for Real Estate Appraisal [78]	Price prediction	In-context learning	Stage 1

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Table 2. Representative domain-specific applications in Nascent-diffusion industries.

Paper	Task	Technique	Stage
Government			
Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents [84]	Social simulation	Simulation MAS	Stage 1
Generative Agent Simulations of 1,000 People [85]	AI for social impact / social simulation	Simulation MAS	Stage 1
GraphCompliance [82]	Compliance	KG	Stage 1–2
PrivComp-KG [83]	Compliance	RAG + KG	Stage 1–2
LLM Based Multi-Agent Generation of Semi-structured Documents [87]	Semi-structured document generation	MAS	Stage 1–2
Agents4Gov [88]	Browser use	Single agent	Stage 1–2

6. Growing-Diffusion Industries

6.1. Trade: Controlled Execution in Compliance-Heavy Workflows

Trade operations span multiple stages, including international commerce, customs clearance, logistics, and compliance. Core information is carried through documents such as invoices, packing lists, and bills of lading, which circulate across customs systems, ERP platforms, and regulatory infrastructures. Unlike content-generation scenarios, trade environments are highly standardized and low–fault-tolerant. Systems must therefore not only generate text but also extract structured fields, provide traceable evidence, and safely interface with downstream processes such as declaration, risk screening, and inspection handling. With the rise of digital trade, AI systems have further expanded toward operational tasks such as inventory management, procurement, and supply-chain coordination, requiring participation in business processes beyond document understanding.

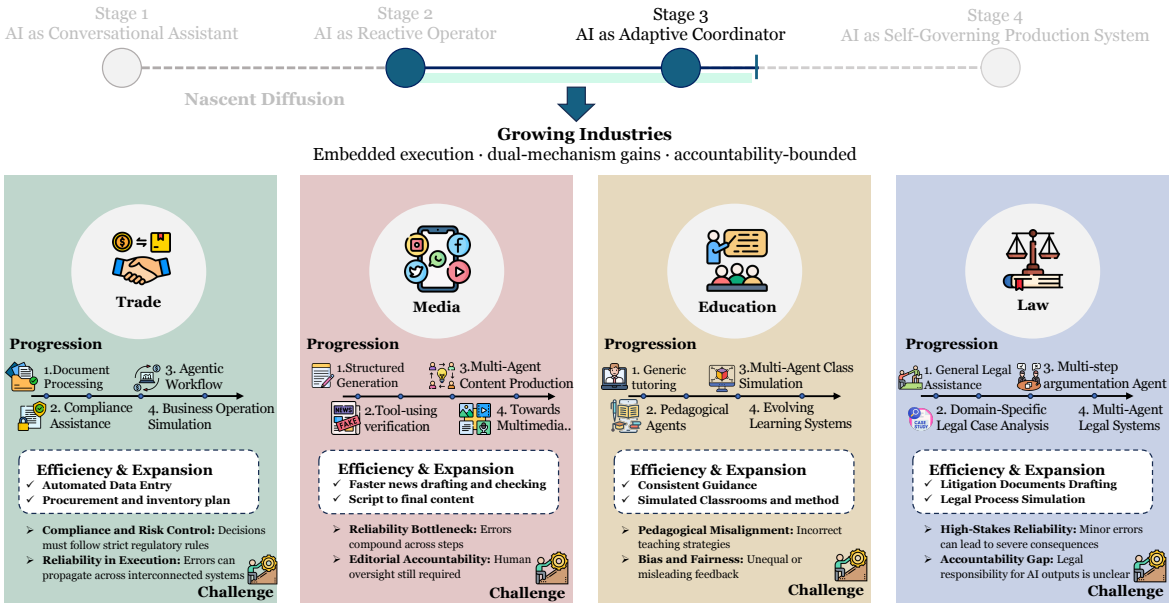


Figure 6. Growing Industries at Stages 2–3 of AI Adoption. These sectors are characterized by embedded execution, dual-mechanism gains in both efficiency and capability expansion, and accountability-bounded deployment. For each industry, the figure traces a four-step progression from basic task assistance toward multi-agent and simulation-based workflows, identifies efficiency and expansion gains, and highlights the key challenges—including reliability bottlenecks, editorial and pedagogical accountability, bias and fairness concerns, and high-stakes responsibility gaps—that constrain progression toward deeper delegation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Document Processing as the First Productivity Layer

Early systems were centered on OCR and intelligent document processing (IDP), transforming unstructured trade documents into structured data for entry, reconciliation, and validation [89]. With the introduction of LLMs, systems evolved toward assisted decision-making with limited semantic understanding and evidence linkage. Multimodal document understanding models integrate text, layout, and visual features to support clause localization, field relation modeling, and cross-field consistency checking, reducing reliance on OCR text alone [90]. Meanwhile, Retrieval-Augmented Generation (RAG) retrieves regulations and historical cases to provide traceable evidence, mitigating hallucination and knowledge drift. In practice, such systems remain positioned as expert-assistance tools in low–fault-tolerance environments. WCO discussions [91] show that AI-assisted classification reduces routine workload while supporting human review through confidence scores and candidate outputs. Industry frameworks increasingly integrate document processing, risk assessment, and classification into unified workflows spanning pre-declaration, clearance, and post-audit stages [92]. However, LLM outputs are typically constrained by evidence retrieval, rule validation, and human review, and remain limited in performing responsibility-bearing operations such as compliance decisions or system submissions [93].

Tool-Orchestrated Workflows for Trade Operations

AI systems are evolving toward agentic workflows that participate in business execution. Tasks such as retrieval, calculation, form filling, submission, and notification are decomposed into traceable tool-invocation chains forming closed-loop workflows [94]. Within this architecture, LLMs handle task understanding and orchestration, while high-risk decisions—such as tariff calculation and sanctions screening—are executed by deterministic modules. This ensures auditability and accountability. Examples include AWS supply-chain systems, where multiple agents coordinate order queries, inventory checks, and rule-based decisions [95]. In customs scenarios, IMF guidance emphasizes that GenAI should support explainable risk analysis while adhering to strict governance frameworks, with RAG as a core component [96].

Why Compliance Responsibility Limits Full Automation

Recent benchmarks suggest improving agent capabilities in long-horizon trade operations. Vending-Bench [32] and its extensions evaluate inventory management, pricing, and profitability over extended horizons, while Anthropic's Project Vend [97] demonstrates early automation in real environments. Multi-agent supply-chain simulations further show efficiency gains in coordination and cost reduction. However, these experiments also reveal instability in strategy and risk control, such as poor inventory management or irrational pricing. Consequently, real-world deployment remains constrained by rule engines, evidence retrieval, and human oversight to mitigate economic and compliance risks.

Overall, AI for Trade is transitioning from document automation to workflow orchestration. While LLM agents can execute multi-step tasks and participate in business processes under controlled settings, real-world systems still require human supervision due to regulatory and financial risks. Current systems are therefore positioned between Stage 2 and Stage 3, with a substantial gap remaining toward unrestricted trade automation.

6.2. Media: Workflow Acceleration under Editorial Accountability

Media production involves multiple linked stages, including reporting, verification, editing, packaging, and distribution. Unlike emerging domains that remain focused on isolated automation, recent work in media increasingly explores systems capable of producing long-form, structured outputs such as full news reports, complete films, or non-linear editing sequences. Research in this field has branched into three distinct paradigms: static workflows for structured generation, autonomous single agents for subtask completion, and multi-agent ecosystems for complex production management.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Structured Generation for Reporting and Content Production

One line of research focuses on utilizing large language models as modular processors within fixed, human-defined workflows, especially in text-based domains. For instance, news generation is often decomposed into fixed stages (e.g. retrieval, semantic scoring, summarization) to produce high-quality long-form reports [98]. Similarly, fact-checking systems like SANCTUARY [99] enforce a strict pipeline of claim generation, evidence retrieval, and ordered reasoning to ensure factual accuracy. In media analysis, the JOA-ICL framework [100] first splits and evaluates structural segments of a full article, and aggregates them to infer the overall position, to enable more accurate article-level stance detection

Agentic Verification, Editing, and Simulation

A second line of research equips agents with the autonomy to interact with external environments, such as search engines or knowledge bases, addressing the need for real-time accuracy and dynamic investigation. This approach is prominent in automated fact-checking, where static retrieval is insufficient. Agents are equipped with various tools, such as web search and source credibility assessment mechanisms, to verify dynamic and complex factual claims [101]. As these autonomous systems become more sophisticated, evaluating their reliability is increasingly critical. To this end, unified frameworks like OpenFactCheck [102] provide comprehensive evaluation pipelines, allowing researchers to customize fact-checking agents and systematically benchmark their verification efficacy against standardized datasets.

Why Editorial Judgment Remains Difficult to Delegate

Multi-agent systems have emerged as a powerful paradigm for organizing complex media generation and evaluation. Within this paradigm, research primarily follows two complementary architectural logics: feedback-driven collaboration which refines outputs through repeated feedback loops between reviewer agents and generator agents, and role-based collaboration where multi-step tasks are completed through role-based division and joint contribution among all agents.

Feedback-driven Collaboration In this paradigm, agents simulate user or expert responses to provide iterative feedback for generator agents or human writers to optimize content quality. In audiovisual media, video editing is modeled as an adversarial dialogue between an editor and a critic agent, allowing for continuous self-correction based on aesthetic feedback [103]. In text-based media, the JRE-L framework [104] iteratively generates and refines the generated news through simulated editorial and reader feedback loops, while the MADES framework deploys memory-augmented agents to debate perspectives and construct tailored supplementary contexts to enhance news understanding [105]. Beyond content creation, this evaluative simulation also extends to media distribution, where personalized recommendation policies are optimized via simulated user interactions, eliminating the need for real-world A/B testing [106,107].

Role-based Coordination In this paradigm, agents collaborate through hierarchical, role-based task decomposition to divide complex tasks into executable sub-tasks. For instance, in audiovisual media generation, approaches like FilmAgent [108] and Kubrick [109] assign agents to specific film crew roles to generate and refine structured scripts and 3D rendering commands for production engines like Unity3D or Blender. Extending these preliminary movie generation approaches, MovieAgent [110] first explores automated long-form video generation by introducing multi-agent Chain of Thought (CoT) planning, while AutoDirector [111] integrates multi-sensory elements such as long takes, special effects, soundtracks, voice dubbing, and lip-synchronization into automatic movie generation, which further enriches the expressiveness of the generated movies.

Collectively, media-oriented LLM agents remain between Stage 2 and the early phase of Stage 3. They can already support multi-step planning, role specialization, and bounded workflow coordination across journalism, recommendation, and audiovisual production, but their autonomy remains limited in real-world use. At present, these systems function primarily as sophisticated tools for

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

accelerating specific sub-tasks, while final judgment, quality control, and release decisions remain under human supervision

6.3. Education: Scalable Personalization under Pedagogical Governance

In education, AI agents are required to do more than retrieve facts: they must understand how to teach, guide students step-by-step, and support knowledge transmission through ongoing interaction, while also meeting stricter pedagogical and ethical requirements because of their potential influence on students' development and values. Accordingly, the following discussion unfolds across three dimensions: the transition from general-purpose models to specialized pedagogical agents, the enhancement of educational interaction through multi-agent architectures, and the evolving assessment and governance of agent autonomy.

Tutoring and Feedback as the Initial Productivity Gains

Early applications of generative AI in education often relied on general-purpose conversational models as virtual tutors [112]. However, the educational domain presents unique and demanding challenges. Beyond information retrieval, effective teaching requires pedagogical intuition, goal decomposition, factuality, and socio-emotional support. General models often struggled with these domain-specific difficulties, frequently providing generic feedback or directly giving away answers rather than appropriately guiding the student. To address these fundamental gaps, the field shifted towards the development of specialized educational models, such as EduChat [113], Google LearnLM [114], and Pedagogy-R1 [115]. By training extensively on pedagogical texts and real-world teaching interactions, these specialized agents are better equipped to handle nuanced tasks like Socratic questioning and complex grading. This shift allows human educators to focus on personalized interventions at a higher-level while verifying the accuracy, tone, and pedagogical alignment of AI-generated feedback [116,117].

Multi-Agent Interaction for Adaptive Learning Workflows

Beyond improving single models for discrete actions, researchers are leveraging multi-agent architectures to facilitate goal-oriented, multi-step reasoning. A primary application is simulation for safe training and strategy evaluation. For example, SocraticLM simulates complete teaching workflows to generate fine-tuning data [118], while Agent4Edu and Classroom Simulacra model virtual students to test pedagogy without human subjects [119,120]. To maintain output quality during extended conversations, systems such as AutoFeedback introduce inspector agents that improve feedback accuracy and reduce common quality issues [121]. Furthermore, the increasing complexity of agent interactions requires more specialized evaluation frameworks. To address this specific demand, MathTutorBench introduces a reward mechanism that penalizes providing direct answers to encourage the model to offer effective guidance [122]. Collectively, these multi-agent advances enable scalable, personalized learning pathways [123,124] and drive the transition of educational agents towards the Stage 3 (AI-as-Adaptive-Coordinator) paradigm.

Why Teaching Autonomy Requires Value Alignment

Educational agent deployment diverges from purely task-oriented systems by necessitating strict alignment with pedagogical paradigms, long-term learning objectives, and human-centric values. Consequently, research increasingly focuses on designing pedagogically aligned and trustworthy architectures. This involves integrating instructional frameworks—such as Socratic scaffolding—into agent workflows while mitigating critical risks like answer leakage and bias amplification [125,126]. Future advancements demand the co-design of cognitive reasoning, interaction protocols, and ethical governance to ensure scalable automation remains anchored in fundamental educational principles [123].

Overall, AI in education is moving from general conversational assistance toward more specialized and interactive support embedded in teaching and learning workflows. Current systems increasingly automate routine pedagogical tasks and enable more adaptive feedback, simulation, and personalized interaction, showing a gradual transition from Stage 2 toward early Stage 3. However, their role

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

remains that of a supervised pedagogical assistant rather than an autonomous substitute for teachers, since real-world deployment is still constrained by the need for pedagogical alignment, value-sensitive guidance, and sustained human oversight.

6.4. Law: Procedural Automation under Professional Responsibility

Legal work is highly reasoning-intensive, procedurally structured, and closely tied to professional responsibility and institutional accountability. Early legal intelligence systems operated in a Conversational Assistant mode, where AI supported retrieval and drafting but remained fully dependent on user control. As legal workflows require reasoning across multiple procedural steps and related laws, systems began to incorporate structured reasoning and external knowledge grounding, reducing cognitive burden but still requiring human supervision. Recent developments in multi-agent coordination and procedural automation indicate a gradual shift toward bounded autonomy. This progression reflects not only technical refinement but also a change in how legal work is organized and delegated between humans and machines.

Legal Retrieval and Drafting as Efficiency Gains

Early work relied on adapting general-purpose language models to legal text. BERT-based models were mainly used for extraction and classification. Later, instruction-tuned generative models enabled drafting and questioning [127]. However, these general models showed clear limitations in legal reasoning and raised concerns about privacy and deployment control. The research focus then moved to training models directly on large-scale legal corpora. Instead of treating legal tasks as simple downstream adaptation, researchers aimed to encode domain knowledge into the model itself. Models such as LawGPT, Lawyer LLaMA, and ChatLaw were developed under this direction [128–130]. SaulLM adopted continual pre-training to maintain general language ability while incorporating legal knowledge [131]. Alignment was also treated more seriously in legal settings. In InternLM-Law, reinforcement learning from human feedback was applied not only to improve output quality, but to enforce professional standards and reduce the risk of generating improper legal strategies [132].

Multi-Step Reasoning and Tool Use in Legal Workflows

Fluency does not guarantee factual accuracy [133]. To reduce hallucination, researchers moved from relying only on parametric memory to using retrieval-augmented generation [134]. This approach allows models to access external legal documents during inference. However, standard vector retrieval is often insufficient for precise legal terminology and complex statute relations. To improve retrieval quality, some systems integrate knowledge graphs to capture structured relations between legal concepts. For example, legal knowledge graph enhanced LLM frameworks combine structured statute relations with language modeling to support multi-step reasoning across related provisions [135]. This allows the model to trace logical connections between statutes instead of relying only on surface similarity. At the same time, structured reasoning methods are introduced to improve logical consistency in complex cases [136]. Frameworks such as Disc-LawLLM apply explicit reasoning formats, including syllogistic reasoning and the IRAC (Issue, Rule, Application, and Conclusion) structure, to make the inference process more transparent and grounded in legal rules [137]. Single law-agents are continuously evolving by integrating instruction decomposition, multimodal visual perception, and advanced reward optimization algorithms like GRPO to reinforce logical reasoning. Despite these advances, single-agent systems remain bounded within legal workflows, as they can support retrieval, reasoning, and drafting but still require human initiation, verification, and approval for consequential legal work.

Why Legal Agency Remains Professionally Bounded

As systems began to tackle more complex legal tasks, researchers began to adopt multi-agent architectures. These systems reflect the division of labor found in real legal practice. Early designs focused on task decomposition. Legal workflows were divided into stages such as intake, retrieval,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

reasoning, and drafting, and each stage was assigned to a specific agent [138]. Multi-agent systems also improve reasoning through structured interaction. In the transactional domain, PAKTON ensures traceable grounding by using a collaborative multi-agent workflow for contract review and question answering [139]. Similarly, LAW uses an orchestrator to connect multiple text agents and domain tools to collaboratively handle complex contract queries like calculating termination dates [140]. Beyond task support, some frameworks explore legal simulation. AgentCourt models trial procedures to study how judicial decisions evolve under structured interaction [141]. Law in Silico constructs simulated digital societies populated by legal agents to evaluate how statutes may produce unintended effects before real-world implementation [142]. These systems move from assistive reasoning toward structured institutional modeling, enabling the analysis of legal processes at a systemic level.

Taken together, legal AI is moving gradually from Stage 2 toward Stage 3. Current systems can decompose goals, coordinate tools, and complete parts of legal workflows through multi-step reasoning, thereby strengthening the *Efficiency Mechanism* mainly through procedural automation and reduced coordination costs. However, their autonomy remains bounded, and human professionals must remain on the loop to connect abstract legal objectives with concrete tasks. Further progress depends on improving citation reliability, robustness under prompt variation, and clearer standards for legal alignment and delegated agency.

6.5. Agentic AI as a Bounded Executor in Growing-Diffusion Domains

Across trade, media, education, and law, Agentic AI is increasingly functioning as a *bounded executor* embedded in real production workflows. As a new productive instrument, its role in these domains is to absorb procedural cognitive labor across multiple intermediate stages of production, including scenario simulation, partial workflow management, iterative revision, and cross-system orchestration under human supervision. Productivity gains in these domains arise mainly through the *Efficiency Mechanism*, with Agentic AI reducing the time, coordination, and verification costs of complex intermediate tasks. A more limited *Expansion Mechanism* is also emerging, enabling broader workflow coverage and larger-scale coordination or simulation.

What limits further progress is the difficulty of making execution both reliable and accountable under the constraints of each domain. Trade requires compliance that can be checked, media requires factual accuracy and editorial oversight, education requires alignment with teaching goals, and law requires clear reasoning supported by traceable evidence. The key bottleneck is that Agentic AI can participate in workflows, but still cannot fully take on responsibility for outcomes. It can already handle many structured substeps, yet it is still weaker at maintaining consistency over long tasks, responding to exceptions, and operating safely when errors carry serious consequences. The main safety risks are also fairly clear, including hallucinated content, unstable multi-step reasoning, bias in sensitive contexts, privacy and intellectual-property leakage, and over-reliance by human users. Taken together, Agentic AI in growing-diffusion domains is best understood as a workflow-embedded productive instrument that raises productivity by taking over increasingly broad ranges of intermediate tasks and by deepening execution across more complex stages of production, while leaving important execution and final judgment in human hands.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Table 3. Representative domain-specific applications in Growing-diffusion industries

Paper	Task	Technique	Stage
Trade			
Azure Microsoft Foundry Document Intelligence [89]	Information extraction	OCR / IDP	Stage 1
Layoutlmv2 [90]	Document understanding	Pretrain & SFT	Stage 1
Mely.ai [94]	Workflow execution	Tool orchestration	Stage 2
AWS supply-chain systems [95]	Supply-chain tasks	MAS	Stage 2–3
GenAI compliance and risk-support systems [96]	Risk analysis	RAG	Stage 1–2
Vending-Bench [32]	Business simulation	Long-horizon benchmark	Stage 2–3
Project Vend [97]	Business automation	Agent-based Simulation	Stage 2–3
Education			
EduChat [113]	Intelligent tutoring	LLM chatbot	Stage 1
LearnLM [114]	Tutoring	Instruction tuning	Stage 1
Pedagogy-R1 [115]	Educational reasoning	Reasoning post-training	Stage 1
SocraticLM [118]	Personalized tutoring	Socratic prompting	Stage 2
LLM-based Automated Grading with Human-in-the-Loop [116]	Automated grading	HITL	Stage 2
AutoFeedback [121]	Automatic feedback	MAS	Stage 2
Agent4Edu [119]	Learner simulation	Generative agents	Stage 2–3
Classroom Simulacra [**?**]	Behavior simulation	Generative agents	Stage 2–3
Law			
LawGPT [128]	Legal drafting	Pretrain & SFT	Stage 1
Lawyer LLaMA [129]	Legal consultation	Pretrain & SFT & Retrieval	Stage 1
InternLM-Law [132]	Legal QA & case analysis	Two-stage SFT	Stage 1
Disc-LawLLM [137]	Legal services & reasoning	SFT & Retrieval	Stage 1
SaulLM [131]	Legal text generation	Pretrain & SFT & DPO	Stage 1
ChatLaw [130]	Legal consultation	MoE & MAS & KG	Stage 2
L-MARS [138]	Legal QA	MAS & Agentic Search	Stage 2
PAKTON [139]	Contract QA & review	MAS & RAG	Stage 2–3
AgentCourt [141]	Judicial decision-making	MAS & Retrieval	Stage 2–3
LAW [140]	Contract query answering	MAS & Tool Orchestration	Stage 3
Law in Silico [142]	Legal society simulation	Agent-based Simulation	Stage 3–4

7. Widespread-Diffusion Industries

7.1. Information Technology: Executable Feedback and Autonomous Software Workflows

As a domain built around highly structured digital artifacts such as code, repositories, and execution logs, information technology provides especially favorable conditions for Agentic AI. Tasks are modular, feedback is often executable, and intermediate outputs can be checked and revised within relatively standardized environments. For this reason, information technology shows one of the clearest and most continuous trajectories from model-level code generation to workflow-level autonomous execution. This evolution can be understood in three stages: post-training to improve single-round code generation, agentic frameworks to manage complex software workflows, and agentic reinforcement learning to strengthen performance in real-world engineering environments.

Code Generation as the Initial Automation Layer

Early work in this domain focused mainly on post-training code large language models to improve single-round generation quality, especially instruction following and local program correctness. Supervised fine-tuning approaches such as WizardCoder [143], together with reinforcement learning methods that incorporate execution feedback such as CodeRL [144], PPOCoder [145], StepCoder [146], and DeepSeek-Coder-V2 [147] demonstrated substantial gains on standalone code generation tasks. However, this line of work largely treated programming as a single-step prediction problem, making it difficult to scale to advanced coding challenges which require workflow orchestration, multi-perspective testing and iterative debugging.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

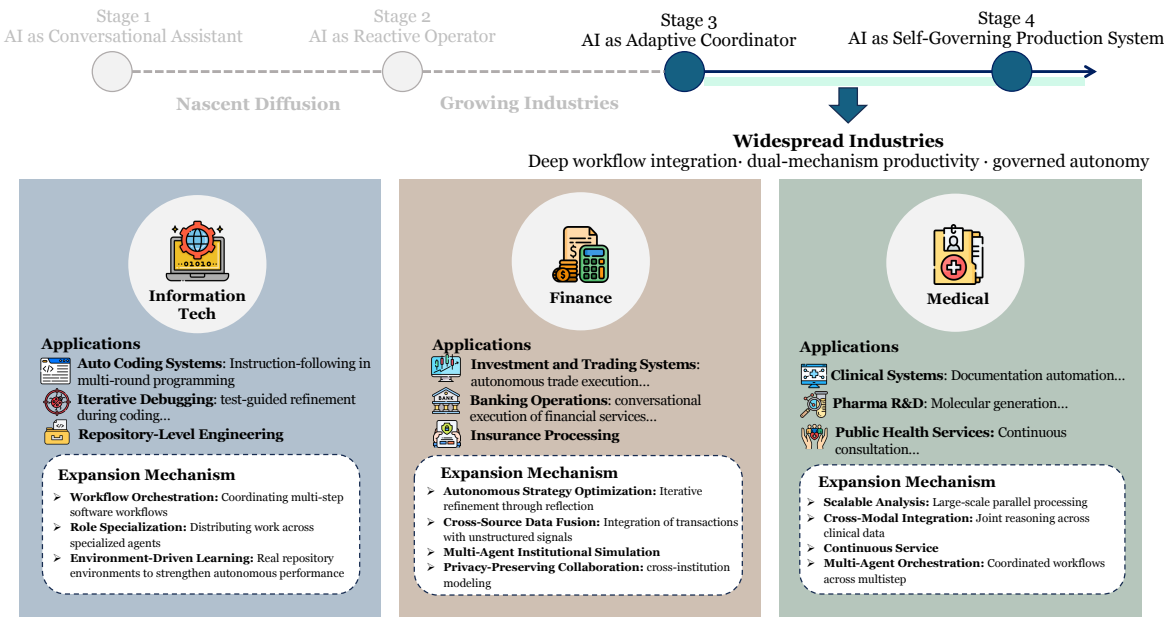


Figure 7. Widespread Industries at Stage 3 and early Stage 4 traits in constrained settings. These sectors are characterized by deep workflow integration, dual-mechanism productivity gains, and governed autonomy. For each industry, the figure identifies representative applications, spanning auto coding, autonomous trading, and clinical documentation, alongside core expansion mechanisms such as workflow orchestration, role specialization, cross-source data fusion, and multi-agent orchestration across complex multistep tasks.

Repository-Level Agents for Iterative Engineering Work

To overcome the limits of single-round code generation, subsequent work shifted from improving isolated outputs to structuring programming as a multi-step workflow. A first step in this direction came from self-debugging methods, which showed that language models could iteratively detect and correct errors through designed prompts or self-generated tests [148–150]. Meanwhile, fixed-pipeline frameworks such as CodePlan [151] showed that manually orchestrated workflows enable LLMs to tackle complex, repository-level coding tasks. Building on this insight, multi-agent systems such as ChatDev [152] introduced explicit role decomposition and standardized operating procedures, recasting software development as a structured pipeline of coordinated subtasks. Expanding this multi-agent paradigm beyond software engineering, MapCoder [153] assigns specialized roles like example retrieval, planning, code generation, and debugging to different agents to solve complex algorithm problems. Together, these frameworks marked a shift from code generation toward bounded workflow execution, but their effectiveness still depended heavily on hand-designed processes and fixed coordination structures

Why Software Becomes the Frontier of Agentic Productivity

As agent-based approaches became more complex, the literature increasingly recognized that procedural orchestration alone could not fully compensate for model limitations in long-horizon reasoning. Real software environments involve changing repository states, hidden dependencies, and delayed feedback, which make fixed workflows brittle. Recent research has therefore returned to strengthening agents' internal reasoning and generalization abilities through reinforcement learning and direct interaction with executable environments. Several works established execution and feedback environments grounded in real-world code repositories, including SWE-Gym [154], AgentGen [155], and R2E-Gym [156], which together form key infrastructure for training and evaluating software engineering agents. Leveraging these environments, DeepSWE [157] demonstrated that large-scale reinforcement learning can significantly improve autonomous exploration and bug-fixing performance under complex dependency structures. RepoForge [158] further advanced this direction by optimizing the end-to-end execution and data pipeline, enabling higher-throughput training at scale. This stage

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

marked a shift from manually designed workflow scaffolds toward agents that can adapt more effectively within real software engineering loops.

Taken together, existing research in information technology shows that AI coding systems have evolved from local developer assistance toward increasingly autonomous execution in software engineering workflows. Agents can already handle goals that are clearly bounded in scope yet require complex, multi-step reasoning, such as resolving real-world GitHub issues through iterative planning, file editing, environment interaction, and testing. In this sense, information technology provides one of the clearest domains in which Agentic AI has moved beyond passive support and toward sustained workflow execution. Based on these capabilities, current systems are best understood as operating at Stage 3 and showing early Stage 4 traits in constrained settings. At the same time, their broader deployment remains constrained by long-horizon reliability and the risks of failure in real development environments.

7.2. Finance: Productivity Amplification under Risk Amplification

As a high-risk domain characterized by data intensity, complex decision logic, and strict regulation, the financial industry requires agents with strong domain specialization, reliable numerical reasoning, and robust compliance awareness. Early systems mainly served as auxiliary tools for text processing and information retrieval, but recent advances have empowered financial agents to participate in core processes such as investment research, trading, and risk management. This pattern indicates that finance can amplify productivity through agentic execution, but it can also amplify risk when autonomous decisions interact with markets, compliance systems, and customer assets.

Financial Reasoning and Research Automation

To address the inherent gaps of general LLMs in financial scenarios, a critical step is to adapt them into specialized financial foundation models with enhanced reasoning and strict compliance. For example, researchers have explored domain-specific pre-training on financial corpora [159], hybrid paradigms that combine LLMs with traditional predictive models [160], and reinforcement learning-based post-training to improve reasoning stability [161]. Compliance capabilities are also integrated into model inference through supervised fine-tuning alongside external post-hoc checks [162,163]. However, these efforts mainly improve reasoning or compliance in isolation, rather than holistic trade-offs among accuracy, traceability, and efficiency in real-world deployment, leaving a clear gap between research progress and practical demands [164,165]. To bridge this gap and fully leverage the improved foundational capabilities, financial AI is evolving toward an agentic stage of proactive task execution, where agents autonomously plan workflows, invoke specialized tools, and iteratively refine strategies around business objectives.

Agentic Execution across Trading, Banking, and Insurance

In practice, current systems increasingly adopt both single-agent specialization and multi-agent collaboration across financial sub-domains.

Securities and Asset Management: In investment research, multi-agent collaboration has become a dominant paradigm. Systems such as FinRobot decompose research workflows into sub-tasks handled by specialized agents, while debate- or verification-based frameworks improve report reliability [166, 167]. Some approaches further integrate code agents and vision models to automate data acquisition and visualization pipelines [168]. Beyond research generation, agents are increasingly applied to trade execution. For example, FinAgent autonomously invokes technical analysis tools and iteratively refines strategies through reflection and memory mechanisms [169]. Multi-agent trading frameworks further simulate institutional decision structures to improve the robustness of strategies, while some systems introduce internal competitive and ranking mechanisms among agents to filter noisy signals in dynamic markets [170,171]. Additionally, reinforcement-learning-based systems such as QuantAgent enable adaptive strategy optimization to capture excess returns in simulated markets [172].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Banking Sector: In banking, intelligent agents are increasingly being integrated into core operational processes. In risk management, LLM-based agents combine structured transaction data with unstructured information such as market signals and public sentiment to support adaptive risk assessment and real-time decision-making [173,174]. Meanwhile, conversational execution agents are emerging as intelligent interfaces for banking systems. Frameworks such as the Ryt AI Banking Agent enable users to perform account operations, payments, and service inquiries through natural language by invoking internal banking APIs [175]. Together, these systems illustrate the transition of banking agents from isolated task automation toward integrated workflow execution.

Insurance Sector: In insurance services, intelligent agents are increasingly used for automated claims processing and fraud detection. Multimodal systems combining computer vision and LLM reasoning can jointly analyze policy documents and damage images to support automated claim assessment [176]. This multimodal fusion also improves fraud detection by integrating visual, textual, and structured claims data [177]. When fraud detection requires data from multiple institutions, but centralized sharing is restricted, federated learning provides a privacy-preserving framework for collaborative modeling across institutions [178].

Why Financial Autonomy Must Remain Governable

Despite rapid progress in agentic financial systems, real-world deployment remains fundamentally constrained by the intrinsic properties of financial markets, including high uncertainty, strong interdependence, and strict regulatory requirements. As a result, current research is increasingly shifting from purely improving task performance to building robust and governable agent architectures, including the integration of risk-aware reasoning, explicit constraint modeling, and compliance-aligned decision processes directly into agent workflows. In parallel, there is a growing emphasis on designing feedback-rich environments and evaluation protocols that reflect real-world financial dynamics, enabling agents to learn and be assessed under more realistic market conditions.

7.3. Medicine: Auditable Human-AI Collaboration in High-Stakes Care

Medical Foundations and Clinical Reasoning Alignment

Early medical AI research focused on integrating the linguistic reasoning abilities of general large language models with specialized medical knowledge. While general models exhibit strong language capabilities, they lack sufficient clinical knowledge coverage and rigor in medical reasoning. To address this gap, early efforts relied on large-scale medical knowledge injection, transforming general models into domain-specific medical foundation models. For instance, PMC-LLaMA [179] integrates biomedical literature and textbooks with medical instruction tuning to enable domain-specific analysis and question answering. Beyond text, generative models have also been extended to biomedical tasks. The PanGu Drug Model [180], trained on large-scale molecular data, supports molecular generation and property prediction, demonstrating that generative capabilities can transfer to drug discovery workflows. As the knowledge base matured, research shifted toward capability alignment. Supervised fine-tuning (SFT) transformed static knowledge into clinical reasoning abilities, as seen in Huatuo-26M [181] and BenTsao [182]. Further work improved interpretability of reasoning, such as ClinicalGPT-R1 [183]. Meanwhile, reinforcement learning and verifiable alignment were introduced to enforce medical safety and reduce bias. For example, HuatuoGPT-o1 [184] designs reward mechanisms for reasoning quality, while approaches inspired by Constitutional AI [185] incorporate ethical constraints via RLAIF.

Agentic Workflows for Documentation, Diagnosis, and Discovery

Medical AI is now transitioning from question-answering systems to executable agentic systems. In clinical settings, single-agent systems integrate deeply with workflows. Systems such as Google MedLM [186] and Microsoft Dragon Copilot [187] connect to EHRs, generate structured clinical notes, retrieve evidence, and assist coding. Similarly, transcription systems like Nuance Copilot [188] and AWS HealthScribe [189] automate clinical documentation, significantly reducing physician workload.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

Multi-agent systems further extend capabilities by simulating collaborative medical teams. Agent Hospital [190] enables iterative learning through simulated cases, while MATEC [191] coordinates specialized agents in critical care scenarios. MedAgent-Pro [192] introduces evidence-based workflows combining planning, tool use, and decision aggregation, reflecting a shift toward controllable and auditable execution. Beyond hospitals, agentization extends to pharmaceutical and public domains. Prompt-to-Pill [193] integrates drug discovery pipelines into end-to-end multi-agent systems, while virtual laboratory platforms [194] enable large-scale parallel hypothesis testing. On the public side, systems such as ChatGPT Health [195] provide continuous health consultation, supporting users in understanding symptoms and navigating healthcare processes. These developments show that AI agents can execute tasks beyond human temporal limits, including large-scale data analysis, real-time transcription, and continuous consultation, becoming increasingly important executors in healthcare systems

Why Healthcare Productivity Depends on Governed Delegation

The development of medical AI has evolved from knowledge injection and capability alignment to agent-centered task execution. Rather than relying solely on scaling, progress increasingly depends on controllability, compliance, and workflow integration. AI improves healthcare through both efficiency mechanisms—automating workflows and assisting decision-making—and expansion mechanisms, such as large-scale molecular screening [193] and continuous public health consultation [195]. Accordingly, evaluation criteria have shifted from response quality to controlled execution and governance. Real-world systems enforce safety through tool constraints, evidence binding, permission control, and human oversight. Most current systems operate at Stage 3 (Adaptive Coordinator), while only early Stage 4 traits appear in constrained settings because safety and regulatory requirements remain binding. Overall, the core value of medical AI lies in transforming complex healthcare processes into scalable and auditable human–AI collaborative systems.

7.4. Agentic AI as Scaled Orchestration in Widespread-Diffusion Domains

Across information technology, finance, and medicine, Agentic AI is an increasingly central *productive instrument* embedded in high-frequency, high-value workflows. In these domains, production already depends heavily on structured digital materials such as code repositories, financial data, transaction records, clinical notes, medical images, and standardized evidence. This makes it easier for agents to take part in continuous workflow orchestration. As a result, Agentic AI is increasingly used as an authorized executor of tasks. It can make detailed plans, retrieve supportive information, and revise intermediate results within rules. Its contribution to productivity works in both width and depth. In width, it now covers a broad set of end-to-end processes, including repository-level code maintenance and debugging, market analysis and claims processing, and clinical documentation, evidence synthesis, and drug discovery support. In depth, delegation has moved to sustained multi-step execution. This makes these domains the clearest cases in which Agentic AI is becoming deeply integrated into production.

This productivity gain is driven by both mechanisms discussed earlier. First, the *Efficiency Mechanism* remains the foundation. Agents reduce the time and verification costs of complex cognitive work by shortening many intermediate steps. Second, the *Expansion Mechanism* is much more visible here than in earlier stages. Agents make it possible to perform tasks at a scale, speed, and level of continuity that human teams alone would struggle to sustain. Examples include parallel code debugging, continuous market monitoring, large-scale evidence synthesis, and always-on public or enterprise services. However, this does not mean that these domains have fully moved into unrestricted autonomy. The main bottleneck is no longer whether agents can produce useful outputs. It is whether they can act reliably under uncertainty while meeting demands for accountability.

This challenge appears in different forms across domains. In Information Technology Industry, the main problem is long-horizon reliability in changing environments with hidden dependencies and real deployment risks. In finance, the key concern is that autonomous action may turn model

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

error into market, compliance, or even systemic risk. In medicine, the barrier is even higher because systems must meet strict requirements for safety, explainability, privacy, and legal responsibility when outputs affect patient care. As deployment becomes more widespread, safety problems also become more concrete. Hallucinated reasoning, prompt injection, unauthorized tool use and privacy leakage are no longer just model-quality issues. They become direct production risks. For this reason, the key task at this stage is not simply to push autonomy further, but to ensure that deeper delegation remains safe, controllable, and accountable.

Table 4. Representative domain-specific applications in Widespread-diffusion industries.

Paper	Task	Technique	Stage
Information Technology			
WizardCoder [143]	Single-round Coding	SFT	Stage 2
CodeRL [144]	Single-round Coding	RL	Stage 2
PPOCoder [145]	Single-round Coding	RL	Stage 2
StepCoder [146]	Single-round Coding	RL	Stage 2
DeepSeek-Coder-V2 [147]	Single-round Coding	RL	Stage 2
Self-Debug [148]	Self-debugging	Agentic Framework	Stage 2–3
Revisit Self-Debugging [149]	Self-debugging	Agentic Framework	Stage 2–3
Large Language Model Guided Self-Debugging Code Generation [150]	Self-debugging	Agentic Framework	Stage 2–3
ChatDev [152]	Repository-level Coding	Agentic Framework	Stage 3
MapCoder [153]	Single-round Coding	Agentic Framework	Stage 3
CodePlan [151]	Repository-level Coding	Agentic Framework	Stage 3
SWE-Gym [154]	Repository-level Coding	Agentic RL	Stage 3–4
AgentGen [155]	Repository-level Coding	Agentic RL	Stage 3–4
R2E-Gym [156]	Repository-level Coding	Agentic RL	Stage 3–4
DeepSWE [157]	Repository-level Coding	Agentic RL	Stage 3–4
RepoForge [158]	Repository-level Coding	Agentic RL	Stage 3–4
Finance			
Alpha-GPT [196]	Alpha mining	Human-AI interaction	Stage 2
Fin-R1 [161]	Financial reasoning	RL	Stage 2
DianJin-R1 [162]	Financial reasoning	RL	Stage 2
FinRobot [166]	Financial analysis	MAS	Stage 3
FinAgent [169]	Trading	Tool augmentation	Stage 3–4
TradingAgents [170]	Trading	MAS	Stage 3
FinDebate [167]	Financial analysis	MAS	Stage 3
FinSight [168]	Financial research	Agentic search	Stage 3–4
RiskLabs [174]	Risk prediction	Multimodal LLM	Stage 2
Credit Risk Meets Large Language Models [160]	Credit scoring	LLM scoring	Stage 2
MMLM-CA [176]	Claims adjudication	Multimodal LLM	Stage 2
Medical			
PMC-LLaMA [179]	Medical QA	SFT	Stage 1
PanGu Drug [180]	Drug discovery	SFT	Stage 1
Huatuo-26M [181]	Clinical reasoning	SFT	Stage 2
BenTsao [182]	Clinical reasoning	SFT	Stage 2
ClinicalGPT-R1 [183]	Medical reasoning	RL	Stage 2
HuatuoGPT-o1 [184]	Medical reasoning	RL	Stage 2
MedLM [186]	Clinical assistance	Agentic RL	Stage 3
Microsoft Dragon Copilot [187]	Clinical assistance	Agentic RL	Stage 3
AWS HealthScribe [189]	Documentation	Agentic RL	Stage 3
Agent Hospital [190]	Diagnosis	MAS	Stage 3
MATEC [191]	Critical care	MAS	Stage 3
MedAgent-Pro [192]	Decision support	MAS	Stage 3
Prompt-to-Pill [193]	Drug discovery	MAS	Stage 3
Virtual Biotech [194]	Experimentation	MAS	Stage 3

8. Cross-Industry Comparison: What Determines Productivity Diffusion?

The cross-industry comparison shows that the uneven diffusion of agentic AI is not simply a matter of model capability. The same technical system can generate substantial productivity gains in one sector while remaining confined to bounded assistance in another. This pattern indicates

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

that productivity diffusion depends on whether a sector can convert agentic capability into reliable workflow execution and accountable delegation.

Digitization of production objects

Agentic AI diffuses more deeply when the objects of production are already digital. Code, documents, structured data, transaction records, and clinical notes can be read, modified, checked, and revised inside software environments. This helps explain why information technology, finance, and parts of medicine show deeper integration than agriculture, manufacturing, and public administration, where physical conditions or institutional procedures limit direct execution.

Feedback verifiability

Productivity gains are more visible when outputs can be verified through executable tests, structured evidence, expert rubrics, or operational metrics. Software has unusually strong feedback because code can be run and tested. By contrast, education, media, government, and healthcare often require qualitative judgment, social context, or professional interpretation, which makes closed-loop autonomy harder to justify.

Error tolerance

The cost of mistakes strongly shapes the depth of delegation. In low-tolerance domains such as finance, healthcare, manufacturing, law, and government, errors can create legal, physical, financial, or public-trust consequences. In these settings, agentic AI may still improve efficiency, but autonomy remains bounded by review, permission control, audit trails, and fallback procedures.

Workflow modularity

Agent orchestration works best when production can be decomposed into explicit subtasks with stable interfaces. Trade documentation, legal drafting, software maintenance, and clinical documentation all contain modular workflow components that agents can execute or coordinate. Where work depends on tacit knowledge, ambiguous goals, or constantly changing environmental conditions, agents remain closer to decision support than delegated execution.

Accountability structure

The hardest barrier is often not whether AI can act, but who remains responsible when it acts. Sectors with clear supervisory structures, auditable outputs, and permission boundaries can delegate more operational responsibility. Sectors where authority, liability, or professional judgment cannot be cleanly transferred keep humans at the center of final decisions. The evidence therefore suggests that productivity diffusion is ultimately governed by the interaction between technical capability and institutional accountability.

9. Discussion and Challenge

9.1. Other Industries

To complement the previous analysis, the following discussion assesses the progress and challenges of AI applications in recent years across several additional sectors (e.g., power and gas, transportation and warehousing, accommodation and catering, and culture and entertainment).

9.1.1. Power and Gas

The power sector is transitioning toward LLM- and agent-driven automated operations, primarily categorized into automated system analysis and intelligent operational control. In system analysis, frameworks like X-GridAgent and GridMind [197,198] automate scientific computing (e.g., optimal power flow) by coupling high-level reasoning with deterministic engineering tools.Expanding on this, recent studies introduce graph and tabular embeddings to help LLMs directly process complex grid constraints for power flow optimization, as demonstrated by PowerGraph-LLM [199]. For operational

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

control, some approaches like Grid-Agent [200] combine semantic planning with domain solvers to coordinate real-time remediation tasks like topology reconfiguration.Specialized models such as GAIA [201] are developed specifically to handle advanced power dispatch operations. Similarly, LLM4DistReconfig [202] utilizes fine-tuned LLMs to rapidly predict optimal network configurations during dynamic topology changes. In summary, while these AI applications reduce operational complexity, challenges in safety, controllability, and real-world deployment remain.

9.1.2. Traffic and Mobility

The transportation sector is transitioning from static, rule-based management to dynamic, LLM-driven autonomous coordination. This evolution addresses operational challenges progressively, from macro-level traffic simulation to micro-level control and routing execution. Initially, evaluating city-scale transportation policies was hindered by the complex manual setup of traffic simulators. To address this, some studies utilize LLM-guided agentic frameworks to automate traffic scenario generation and simulation through natural language, exemplified by AgentSUMO [203] and TrafficSimAgent [204]. To overcome the traditional reliance on rigid heuristics for intersection traffic control, certain approaches employ LLM-based agents equipped with common sense and reasoning capabilities to dynamically optimize signal timings, as demonstrated by LLMLight [205]. Regarding physical fleet execution, solving dynamic Vehicle Routing Problems traditionally demanded extensive human expertise to design routing algorithms. Recent frameworks like VRPAgent [206] and AFL [207] leverage LLMs to autonomously discover and refine heuristic operators for complex fleet coordination. Overall, while these progressive applications highlight the immense potential of AI agents across transportation systems, challenges in real-time latency and multi-agent safety remain.

9.1.3. Accommodation and Catering

The accommodation and catering sectors are transitioning from labor-intensive service models to AI-integrated autonomous operations. To overcome early constraints imposed by manual staff availability and static pricing rules, some studies employ LLM-driven conversational agents to automate complex front-desk booking and inquiry workflows, as demonstrated by recent systems [208]. For backend optimization, certain approaches utilize predictive AI to dynamically forecast demand and adjust pricing, replacing traditional empirical models [209,210]. Moving to the catering domain, to overcome efficiency and consistency bottlenecks in manual multi-dish preparation, researchers develops customized robotic systems to automate physical cooking tasks, such as the dual-arm robot YORI [211]. To further enhance kitchen autonomy, some studies have leveraged large language models to translate natural-language recipes into structured task trees and executable robotic cooking plans for embodied kitchen systems [212,213].

9.1.4. Culture and Entertainment

The cultural and entertainment sectors are transitioning from manual, resource-intensive production to generative AI and agent-driven interactive ecosystems. Traditional manual asset creation carries high temporal and financial costs. In response, multi-agent frameworks such as FilmAgent [108] and MAViS [214], together with modern generative pipelines [215], have emerged to automate scriptwriting, cinematography, and long-sequence video storytelling workflows. Within gaming, to overcome traditional limitations, recent workflows integrate generative AI to accelerate character creation pipelines [216], while LLM-based NPC systems are enabling more interactive and active in-game characters [217,218]. Similarly, AI-enhanced conversational systems are now deployed in public venues like museums to replace static, one-way exhibit interpretations with dynamically enriched visitor experiences [219]. However, alongside these productivity gains, the expansion of generative AI continues to intensify critical debates over copyright, authorship, and creative integrity [220,221].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

9.2. Remaining Challenges on Productivity Diffusion

These challenges suggest that the next stage of AI for Productivity will be determined less by model capability alone than by workflow integration, verification infrastructure, permission design, and institutional accountability.

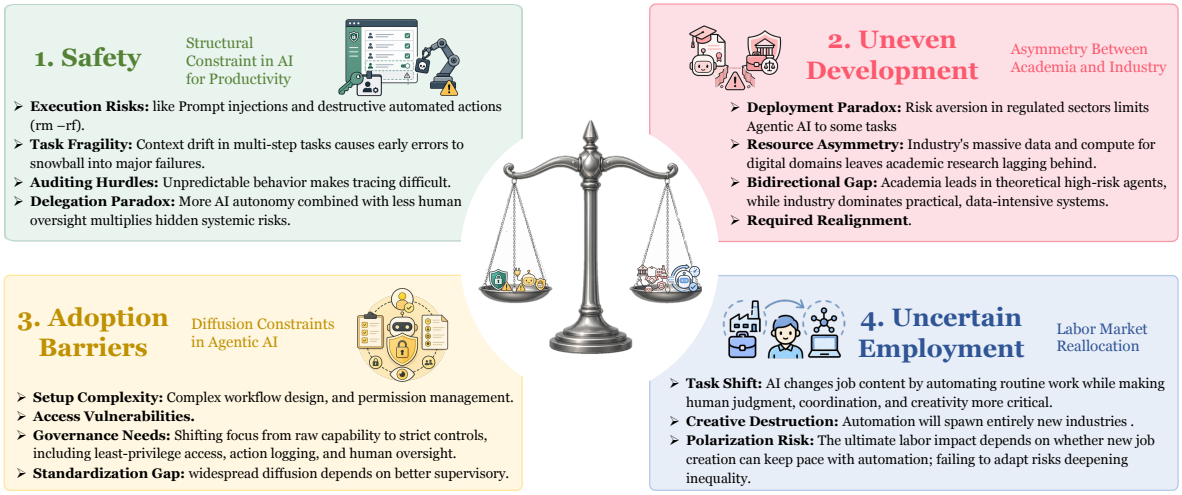


Figure 8. Structural Constraints on Productivity Diffusion. This figure outlines four constraints on the broader diffusion of agentic AI: (1) safety and long-horizon reliability risks from execution errors, task fragility, and reduced oversight; (2) the capability-deployment gap between academia and industry; (3) adoption friction from setup complexity, governance needs, and standardization gaps; and (4) labor reallocation and employment uncertainty from task shifts and labor market polarization. Together, they indicate that the future of agentic AI depends on balanced progress in safety, governance, and socioeconomic adaptation.

9.2.1. Safety and Long-Horizon Reliability

Safety is an essential constraint in AI for productivity. As generative AI and high-permission agentic systems are increasingly integrated into enterprise production workflows, system safety concerns are evolving from traditional issues of software vulnerabilities and access control into a more complex risk structure involving model behavior, data flows, and decision execution. Unlike earlier information systems that primarily functioned as passive tools, AI systems in production environments actively participate in reasoning, decision-making, and even cross-system operations. As a result, risks extend beyond information leakage to higher-level issues such as behavioral deviation and loss of execution control.

In practice, these risks have already manifested in diverse forms. First, at the data level, existing studies and industry reports indicate that a large number of enterprise employees routinely input customer data, financial records, and even proprietary code into external generative models, thereby bypassing established data governance and access control mechanisms and significantly increasing the risk of sensitive information leakage [222,223]. Second, at the model level, prompt injection has been widely recognized as one of the primary attack vectors in generative AI systems. Through carefully crafted inputs, attackers can induce models to deviate from their intended objectives, potentially exposing internal knowledge or performing unauthorized actions [224,225]. Third, at the system level, as agents become deeply integrated with toolchains, high-permission agents may trigger destructive behaviors during automated execution due to policy misalignment or contextual misunderstanding, such as deleting critical data or executing operations that violate compliance constraints. These issues are particularly pronounced in low-tolerance domains such as finance, healthcare, and public administration, where even minor errors can lead to systemic consequences or legal liabilities.

Long-Horizon Reliability as a Structural Source of Risk

A key source of safety risk lies in the structural mismatch between long-horizon task requirements and the capability profile of current models. Real-world production tasks typically take the form

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

of multi-stage, tool-integrated continuous control processes, where correctness is defined at the trajectory level rather than by the accuracy of individual steps [226,227]. However, current LLMs, trained under next-token prediction objectives, lack native mechanisms for modeling long-term state and constraint consistency, making them prone to intention drift, constraint violations, and breakdowns in intermediate reasoning during execution. This structural fragility further manifests as error compounding in long sequences: small early deviations can propagate and amplify by altering subsequent decision contexts, ultimately leading to unpredictable system behavior. Meanwhile, context drift weakens the persistence of critical constraints, causing models to gradually deviate from initial objectives over extended interactions. In high-permission settings, these issues directly translate into safety risks, including unauthorized actions, incorrect decision execution, and unintended data exposure. Therefore, insufficient long-horizon reliability is not only a performance bottleneck but also a foundational constraint on system safety.

Sustainability Constraints and Persistent Security Vulnerabilities

Another class of safety challenges arises from sustainability constraints in the long-term operation of AI productivity systems. First, the non-deterministic behavior of models and ongoing version evolution reduce reproducibility, complicating security auditing, fault tracing, and compliance verification [228,229]. Second, sustained capability improvement relies heavily on external data and environment interaction, while high-quality data—particularly realistic long-horizon task trajectories—remains scarce in practice, thereby limiting robust improvements in complex decision-making scenarios. This constraint on data availability and capability growth makes systems more likely to be deployed or scaled without sufficient validation, leading to the accumulation of latent risks [230]. In addition, continuous deployment environments introduce new attack surfaces, such as prompt injection and toolchain-based attacks, transforming safety from a one-time design concern into an ongoing operational challenge [231]. At the same time, as AI systems take on more responsibilities, human involvement in supervision and review may diminish, weakening the ability to detect anomalous behavior. This dynamic of "increased delegation and reduced oversight" further amplifies systemic safety risks.

9.2.2. The Capability-Deployment Gap between Academia and Industry

Another challenge of the current AI landscape is a bidirectional asymmetry between academic research and industrial reality. While academia often leads in exploring high-risk agents, industry leads in building data-intensive systems.

In high-stakes, regulated sectors such as finance and law, academic research often explores Stage 3 autonomous agents capable of complex portfolio management or legal reasoning. However, real-world deployment lags significantly behind these theoretical capabilities. Industry reports indicate that adoption is restricted to low-risk, auxiliary functions such as customer support and document drafting, rather than core decision-making processes [232]. This cautious approach creates a paradox of limited economic returns. Since organizations limit AI to low-impact tasks to reduce legal and regulatory risk, potential productivity gains are also restricted, which further discourages the deeper integration needed to fully realize the value of autonomous AI systems [233].

In contrast, this dynamic is reversed in digital-native domains such as software engineering. Leveraging large proprietary datasets, expanded GPU resources, and user feedback loops that remain inaccessible to open-source researchers, proprietary coding agents (e.g. Claude Code [234], Codex [235], Cursor [236], OpenClaw [10], WisPaper [237]) consistently outperform academic and open-source systems across a range of benchmarks requiring repository-level reasoning and end-to-end task completion [238]. In this context, the uneven distribution of resources results in slower progress in academic research compared to industrial development.

Consequently, a substantial gap exists between the theoretical potential of AI agents and their practical implementation across industries. Addressing this gap is critical, since uneven progress may leave high-risk industries without effective automation capabilities, while also leading to the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

concentration of core functionalities within closed, proprietary systems. To align these developments, future research efforts necessitate enhanced coordination between academia and industry. Academic research should shift its focus from sole emphasis on performance metrics to prioritizing reliability and auditability for regulated sectors. Additionally, the open research community should develop data-efficient methodologies to maintain parity with industrial-scale systems.

9.2.3. Adoption Friction in Real Organizations

A third challenge is that agentic AI still faces high barriers to broad adoption. Recent systems show that AI can move beyond text-based assistance toward tool use and multi-step task execution, but this has not yet translated into wide deployment across ordinary users and firms [43,239]. The gap reflects not only model limits, but also the practical difficulty of deploying, supervising, and safely governing agents in real settings [240].

These barriers come from several sources. Many agent systems still require complex setup, including tool connection, workflow design, and permission management, which keeps effective use concentrated among technically capable users and organizations [240]. At the same time, stronger execution ability often requires broader access to files, browsers, APIs, and enterprise systems, which makes mistakes more costly and expands the attack surface. This is especially important because agent systems blur the line between data and instructions and remain vulnerable to prompt-injection style attacks [241,242].

Current responses therefore focus less on raw capability alone and more on control, identity, and governance. Recent standards and governance work places growing weight on authentication, authorization, least-privilege access, action logging, and human approval for high-risk actions [243,244]. This suggests that broader diffusion will depend not only on making agents stronger, but also on making them easier to monitor, constrain, and integrate into existing institutions [245].

This pattern is similar to the early development of personal computers. Early PC systems also had high barriers and diffused slowly at first. Wider adoption came later through better interfaces, standardized platforms, and practical applications that made the technology useful to ordinary users and firms [246,247]. Agentic AI may now be in a similar middle stage between early experimentation and wider diffusion. Looking ahead, progress will likely depend on narrow high-value use cases, stronger supervisory interfaces, and clearer standards for permissions, tool use, and auditing [243,248].

9.2.4. Labor Reallocation and Employment Uncertainty

A fourth challenge is that the employment effects of agentic AI remain highly uncertain. Agents may automate some tasks and raise productivity, but they may also change job structure, reshape labor demand, and create adjustment pressures before new employment opportunities fully emerge.

One way to understand this is through the task-based framework [249]. This view treats production as a set of tasks performed by labor, capital, or algorithms, rather than as a fixed set of occupations. Agentic AI therefore matters because it changes which tasks can be automated, which still rely on human labor, and how companies reorganize work internally [41,250,251]. As a result, the labor-market effects of Agentic AI are likely to appear first through changes in job content. Routine and standardized tasks may weaken, while judgment, coordination, and creative tasks become relatively more important.

A broader perspective comes from endogenous growth theory, which emphasizes innovation, knowledge accumulation, and the creation of new products and activities as sources of long-run growth [3,252,253]. In this view, AI affects employment through a process of creative destruction. Some jobs may contract, but new industries and new forms of work may also emerge. The overall outcome therefore depends on whether new task creation, organizational adaptation, and worker reallocation can keep pace with automation. If they do not, AI may deepen polarization and inequality. If they do, AI may raise productivity while also supporting new employment opportunities.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

10. Conclusions and Further Outlook

The central argument of this paper is that Agentic AI should be understood as a new productive instrument that is beginning to reorganize how work is executed, coordinated, and scaled. Across industries, its economic value comes from two closely connected mechanisms. First, it reduces the time and labor required for existing tasks and therefore improves productive **efficiency**. Second, it makes a broader range of economically valuable tasks feasible at acceptable cost and thereby **expands** the scope of production. Yet the diffusion of this new productive instrument remains highly uneven across sectors. In some industries, AI still plays a bounded assistive role. In others, it has already been embedded in multi-step workflows as a supervised executor. In a small number of digital-native domains, it is moving toward more sustained autonomous agency. Taken together, the importance of Agentic AI lies not only in improving efficiency at the task level, but also in its growing capacity to reshape the mode of production itself.

This also means that the impact of Agentic AI increasingly goes beyond benchmarks and cannot be captured by numbers alone. As agents take over more execution, coordination, monitoring, and routine decision-making, production relations themselves are beginning to change. The role of human labor is gradually shifting away from directly completing large numbers of intermediate tasks and toward goal-setting, handling unexpected exceptions, and verifying final outputs. At the organizational level, firms may increasingly redesign workflows around hybrid human and agent systems, reallocate control across layers of execution, and create new divisions of labor among workers, managers, and machine agents. For this reason, the key long-run question is not simply how much output AI can generate, but how it changes who performs which tasks, who controls the workflow, and who remains accountable for the outcome. The future of AI for Productivity therefore depends not only on further gains in capability, but also on progress in reliability, safety, evaluation, and institutional governance.

Looking ahead, **AI for Productivity** can be seen as the first major effect of Agentic AI in a broader developmental trajectory. After directly transforming production, the next step is likely to be **AI for Science**. At this stage, Agentic AI will begin to accelerate scientific discovery itself through literature review, hypothesis generation, simulation, experimentation, and cross-disciplinary knowledge integration. Beyond that lies **AI for AI**. As human capability approaches its limits in some parts of AI development, Agentic AI will increasingly participate in improving AI itself, including data, architectures, harnesses, and evaluation systems. At the macro level, AI for Productivity, AI for Science, and AI for AI may form a mutually reinforcing loop. AI reorganizes production, stronger productive capacity supports faster scientific discovery, and advances in science and engineering feed back into more capable AI systems, which then further reshape production. The deepest impact of Agentic AI may therefore lie not in automating isolated tasks, but in driving the emergence of a new socio-technical production system in which economic activity, knowledge creation, and AI development increasingly coevolve. The central insight is that AI for Productivity is not merely a story of faster task completion. It is a story of delegated agency: as AI systems take over execution, coordination, and monitoring, productivity gains increasingly depend on how organizations redesign workflows, assign responsibility, and govern autonomous action.

References

1. Smith, A. *The wealth of nations [1776]*; Vol. 11937, 1937.
2. Marx, K. *Capital: Critique of Political Economy*; Vol. 1, 1867.
3. Romer, P. Endogenous Technological Change. Working Paper 3210, National Bureau of Economic Research, 1989. https://doi.org/10.3386/w3210.
4. Chen, Z.; Shen, S.; Shen, G.; Zhi, G.; Chen, X.; Lin, Y. Towards Tool Use Alignment of Large Language Models. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y., Eds. Association for Computational Linguistics, 2024, pp. 1382–1400. https://doi.org/10.18653/V1/2024.EMNLP-MAIN.82.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

5. Yan, S.; Yang, X.; Huang, Z.; Nie, E.; Ding, Z.; Li, Z.; Ma, X.; Schütze, H.; Tresp, V.; Ma, Y. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning. *CoRR* **2025**, *abs/2508.19828*, [2508.19828]. https://doi.org/10.48550/ARXIV.2508.19828.
6. Chhikara, P.; Khant, D.; Aryan, S.; Singh, T.; Yadav, D. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413* **2025**.
7. Packer, C.; Fang, V.; Patil, S.G.; Lin, K.; Wooders, S.; Gonzalez, J.E. MemGPT: Towards LLMs as Operating Systems. **2023**.
8. Erdogan, L.E.; Lee, N.; Kim, S.; Moon, S.; Furuta, H.; Anumanchipalli, G.; Keutzer, K.; Gholami, A. Plan-and-Act: Improving Planning of Agents for Long-Horizon Tasks. In Proceedings of the Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025; Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; Zhu, J., Eds. PMLR / OpenReview.net, 2025, Vol. 267, *Proceedings of Machine Learning Research*.
9. Xiong, W.; Song, Y.; Dong, Q.; Zhao, B.; Song, F.; Wang, X.; Li, S. Mpo: Boosting llm agents with meta plan optimization. *arXiv preprint arXiv:2503.02682* **2025**.
10. openclaw. OpenClaw: Your own personal AI assistant. https://github.com/openclaw/openclaw, 2026. GitHub repository, starred 185,000+.
11. Anthropic. Claude code, 2025.
12. Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* **2025**, *68*, 121101.
13. Hao, Y.; Chen, Y.; Zhang, Y.; Fan, C. Large language models can solve real-world planning rigorously with formal verification tools. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 3434–3483.
14. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach (4th Edition)*; Pearson, 2020.
15. OECD. *Measuring Productivity – OECD Manual: Measurement of Aggregate and Industry-level Productivity Growth*; OECD Publishing: Paris, 2001. Defines labor productivity as the ratio of volume measure of output to a measure of input use., https://doi.org/10.1787/9789264194519-en.
16. Solow, R.M. Technical Change and the Aggregate Production Function. *The Review of Economics and Statistics* **1957**, *39*, 312–320.
17. Rosenberg, N. *Inside the Black Box: Technology and Economics*; Cambridge University Press, 1983.
18. Bresnahan, T.F.; Trajtenberg, M. General purpose technologies 'Engines of growth'? *Journal of Econometrics* **1995**, *65*, 83–108. https://doi.org/https://doi.org/10.1016/0304-4076(94)01598-T.
19. Helpman, E. *General purpose technologies and economic growth*; MIT press, 1998.
20. Atkeson, A.; Kehoe, P.J. The Transition to a New Economy After the Second Industrial Revolution. Working Paper 8676, National Bureau of Economic Research, 2001. https://doi.org/10.3386/w8676.
21. David, P. The Dynamo and the Computer: An Historical Perspective On the Modern Productivity Paradox. *American Economic Review* **1990**, *80*, 355–61.
22. Brynjolfsson, E.; McAfee, A. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*; W. W. Norton & Company: New York, 2014.
23. Autor, D.H. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives* **2015**, *29*, 3–30. https://doi.org/10.1257/jep.29.3.3.
24. Acemoglu, D.; Restrepo, P. Artificial Intelligence, Automation and Work. NBER Working Papers 24196, National Bureau of Economic Research, Inc, 2018.
25. Zhou, E.; Xi, Z.; Ma, L.; Zhang, Z.; Dou, S.; Lei, Z.; Wang, G.; Zheng, R.; Yan, H.; Gui, T.; et al. Steering LLMs via Scalable Interactive Oversight. *arXiv preprint arXiv:2602.04210* **2026**.
26. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023.
27. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023; Follmer, S.; Han, J.; Steimle, J.; Riche, N.H., Eds. ACM, 2023, pp. 2:1–2:22. https://doi.org/10.1145/3586183.3606763.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

28. Dao, T.; Fu, D.Y.; Ermon, S.; Rudra, A.; Ré, C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
29. International Data Corporation (IDC). Worldwide Spending on Artificial Intelligence Forecast to Reach $632 Billion in 2028, According to a New IDC Spending Guide. Press Release, 2024. IDC forecasts global artificial intelligence (AI) spending, including generative AI, will more than double by 2028 when it is expected to reach $632 billion with strong growth across industries and technology categories.
30. Singh, M.; Rana, A.; Chintala, N.; Shum, C.; Tseng, S. Generative AI Races Toward $1.3 Trillion in Revenue by 2032, 2024.
31. Patwardhan, T.; Dias, R.; Proehl, E.; Kim, G.; Wang, M.; Watkins, O.; Fishman, S.P.; Aljubeh, M.; Thacker, P.; Fauconnet, L.; et al. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. *CoRR* **2025**, *abs/2510.04374*, [2510.04374]. https://doi.org/10.48550/ARXIV.2510.04374.
32. Backlund, A.; Petersson, L. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. *CoRR* **2025**, *abs/2502.15840*, [2502.15840]. https://doi.org/10.48550/ARXIV.2502.15840.
33. Hu, X.; Xia, J.; Xu, S.; Song, K.; Yuan, Y.; Zhang, G.; Ren, J.; Feng, B.; Lu, L.; Zeng, T.; et al. EcoGym: Evaluating LLMs for Long-Horizon Plan-and-Execute in Interactive Economies, 2026, [arXiv:cs.CL/2602.09514].
34. Vidgen, B.; Fennelly, A.; Pinnix, E.; Mahapatra, C.; Richards, Z.; Bridges, A.; Huang, C.; Hunsberger, B.; Zafar, F.; Foody, B.; et al. The AI Productivity Index (APEX). *CoRR* **2025**, *abs/2509.25721*, [2509.25721]. https://doi.org/10.48550/ARXIV.2509.25721.
35. Mazeika, M.; Gatti, A.; Menghini, C.; Sehwag, U.M.; Singhal, S.; Orlovskiy, Y.; Basart, S.; Sharma, M.; Peskoff, D.; Lau, E.; et al. Remote Labor Index: Measuring AI Automation of Remote Work. *CoRR* **2025**, *abs/2510.26787*, [2510.26787]. https://doi.org/10.48550/ARXIV.2510.26787.
36. Wijk, H.; Lin, T.R.; Becker, J.; Jawhar, S.; Parikh, N.; Broadley, T.; Chan, L.; Chen, M.; Clymer, J.; Dhyani, J.; et al. RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. In Proceedings of the Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025; Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; Zhu, J., Eds. PMLR / OpenReview.net, 2025, Vol. 267, *Proceedings of Machine Learning Research*.
37. Dell'Acqua, F.; McFowland III, E.; Mollick, E. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Working Paper 24-013, Harvard Business School, 2023.
38. Arnon, A. The Projected Impact of Generative AI on Future Productivity Growth, 2025.
39. Bresnahan, T.F.; Brynjolfsson, E.; Hitt, L.M. Information Technology, Workplace Organization and the Demand for Skilled Labor: Firm-Level Evidence. NBER Working Papers 7136, National Bureau of Economic Research, Inc, 1999.
40. Raisch, S.; Krakowski, S. Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review* **2021**, *46*, 192–210. https://doi.org/10.5465/amr.2018.0072.
41. Acemoglu, D. The simple macroeconomics of AI*. *Economic Policy* **2024**, *40*, 13–58, [https://academic.oup.com/economicpolicy/article-pdf/40/121/13/59091970/eiae042.pdf]. https://doi.org/10.1093/epolic/eiae042.
42. Peng, S.; Kalliamvakou, E.; Cihon, P.; Demirer, M. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. *CoRR* **2023**, *abs/2302.06590*, [2302.06590]. https://doi.org/10.48550/ARXIV.2302.06590.
43. Singla, A.; Sukharevsky, A.; Yee, L.; Chui, M.; Hall, B.; Balakrishnan, T. The State of AI in 2025: Agents, Innovation, and Transformation, 2025. McKinsey Global Survey on AI.
44. Kalliamvakou, E. Research: Quantifying GitHub Copilot's Impact on Developer Productivity and Happiness, 2022.
45. Brynjolfsson, E.; Li, D.; Raymond, L. Generative AI at Work*. *The Quarterly Journal of Economics* **2025**, *140*, 889–942, [https://academic.oup.com/qje/article-pdf/140/2/889/61701561/qjae044.pdf]. https://doi.org/10.1093/qje/qjae044.
46. Anthropic. Cowork: Claude Code power for knowledge work | Claude by Anthropic, 2026.
47. Anthropic. Claude Code for Enterprise | Claude by Anthropic, 2026.
48. Anthropic. Customer story | Rakuten, 2026.
49. Anthropic. Customer story | Wiz, 2026.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

50. Anthropic. Customer story | Airtree, 2026.
51. Anthropic. Customer story | Zapier Q&A, 2026.
52. Yang, Q.; Liu, Y.; Li, J.; Bai, J.; Chen, H.; Chen, K.; Duan, T.; Dong, J.; Hu, X.; Jia, Z.; et al. OneMillion-Bench: How Far are Language Agents from Human Experts? *arXiv preprint arXiv:2603.07980* **2026**.
53. Chen, K.; Ren, Y.; Liu, Y.; Hu, X.; Tian, H.; Xie, T.; Liu, F.; Zhang, H.; Liu, H.; Gong, Y.; et al. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *arXiv preprint arXiv:2506.13651* **2025**.
54. Kwa, T.; West, B.; Becker, J.; Deng, A.; Garcia, K.; Hasin, M.; Jawhar, S.; Kinniment, M.; Rush, N.; Von Arx, S.; et al. Measuring AI Ability to Complete Long Software Tasks. *Advances in Neural Information Processing Systems* **2026**, *38*, 92213–92266.
55. Roth, K. Towards Total Recall in Industrial Anomaly Detection. *TIB Data Manager* **2021**. https://doi.org/10.57702/mm9ulkog.
56. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization. *Lecture notes in computer science* **2020**, pp. 475–489. https://doi.org/10.1007/978-3-030-68799-1_35.
57. Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; Steger, C. The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. *International Journal of Computer Vision* **2021**, *129*, 1038–1059. https://doi.org/10.1007/s11263-020-01400-4.
58. Electric, S. Building Sustainability's Digital Future with EcoStruxure™ Resource Advisor Copilot: Schneider Electric's Latest AI Advancement, 2024. Press release on the launch of a conversational AI tool embedded in EcoStruxure Resource Advisor Copilot for sustainability and energy data management.
59. Siemens. Siemens Xcelerator: Scaling Roll-Out of Generative AI with Siemens Industrial Copilot. Press Release, 2024. Press release on Siemens' expansion of generative AI via the Siemens Industrial Copilot connected to the Siemens Xcelerator platform at Hannover Messe 2024.
60. ABB. Helping industries do better with generative AI: ABB launches Genix Copilot with Microsoft. Press Release, 2024. ABB Ability™ Genix Copilot integrates generative AI to boost productivity, efficiency and sustainability in industrial operations in collaboration with Microsoft.
61. Honeywell. Honeywell Unveils Innovative AI Assistant For Industrial Operators In Honeywell Forge Production Intelligence. Press Release, 2025. Honeywell announced a new generative AI assistant embedded in Honeywell Forge Production Intelligence to help industrial operators streamline tasks, automate troubleshooting and enhance production insights.
62. Siemens. Siemens Introduces AI Agents for Industrial Automation. Press Release, 2025. At Automate 2025 in Detroit, Siemens announced an expansion of its industrial AI offerings with advanced AI agents designed to operate autonomously within its Industrial Copilot ecosystem.
63. MarketsandMarkets. Artificial Intelligence in Manufacturing Market worth $155.04 billion by 2030 – Exclusive Report by MarketsandMarkets™. Press Release, 2025. The global AI in manufacturing market is projected to grow from USD 34.18 billion in 2025 to USD 155.04 billion by 2030, registering a CAGR of 35.3 % according to a new report by MarketsandMarkets™.
64. Technavio. Generative AI In Manufacturing Market: On-premises is expected to lead the Deployment segment during 2025-2029. Press Release / Market Report Summary, 2025. The global generative AI in manufacturing market is expected to grow at a CAGR of 37.7 % from 2024 to 2029, driven by the need for accelerated product innovation, operational efficiency and human-AI collaboration in manufacturing environments.
65. Rezayi, S.; Liu, Z.; Wu, Z.; Dhakal, C.; Ge, B.; Zhen, C.; Liu, T.; Li, S. AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition. In Proceedings of the Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022, pp. 5150–5156. https://doi.org/10.24963/ijcai.2022/715.
66. Awais, M.; Alharthi, A.H.S.A.; Kumar, A.; Cholakkal, H.; Anwer, R.M. AgroGPT: Efficient Agricultural Vision-Language Model with Expert Tuning. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025, pp. 5687–5696. https://doi.org/10.1109/WACV61041.2025.00555.
67. Samuel, D.J.; Skarga-Bandurova, I.; Sikolia, D.; Awais, M. AgroLLM: Connecting Farmers and Agricultural Practices through Large Language Models for Enhanced Knowledge Transfer and Practical Application. *CoRR* **2025**, *abs/2503.04788*, [arXiv:cs.CL/2503.04788]. https://doi.org/10.48550/arXiv.2503.04788.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

68. Yang, S.; Liu, Z.; Mayer, W.; Ding, N.; Wang, Y.; Huang, Y.; Wu, P.; Li, W.; Li, L.; Zhang, H.Y.; et al. ShizishanGPT: An Agricultural Large Language Model Integrating Tools and Resources. In Proceedings of the Web Information Systems Engineering – WISE 2024, Singapore, 2025; Vol. 15439, *Lecture Notes in Computer Science*, pp. 284–298. https://doi.org/10.1007/978-981-96-0573-6_21.
69. Zhang, Y.; Fan, Q.; Chen, X.; Li, M.; Zhao, Z.; Li, F.; Guo, L. IPM-AgriGPT: A Large Language Model for Pest and Disease Management with a G-EA Framework and Agricultural Contextual Reasoning. *Mathematics* **2025**, *13*, 566. https://doi.org/10.3390/math13040566.
70. Yang, B.; Zhang, Y.; Feng, L.; Chen, Y.; Zhang, J.; Xu, X.; Aierken, N.; Li, Y.; Chen, Y.; Yang, G.; et al. AgriGPT: a Large Language Model Ecosystem for Agriculture. *CoRR* **2025**, *abs/2508.08632*, [arXiv:cs.CL/2508.08632]. https://doi.org/10.48550/arXiv.2508.08632.
71. Wang, L.; Jin, T.; Yang, J.; Leonardis, A.; Wang, F.; Zheng, F. Agri-LLaVA: Knowledge-Infused Large Multimodal Assistant on Agricultural Pests and Diseases. *CoRR* **2024**, *abs/2412.02158*, [arXiv:cs.CV/2412.02158]. https://doi.org/10.48550/arXiv.2412.02158.
72. Yang, B.; Chen, Y.; Feng, L.; Zhang, Y.; Xu, X.; Zhang, J.; Aierken, N.; Huang, R.; Lin, H.; Ying, Y.; et al. AgriGPT-VL: Agricultural Vision–Language Understanding Suite. OpenReview, 2025. ICLR 2026 withdrawn submission.
73. Yang, B.; Feng, L.; Chen, Y.; Zhang, Y.; Zhang, J.; Xu, X.; Aierken, N.; Li, S. AgriGPT-Omni: A Unified Speech-Vision-Text Framework for Multilingual Agricultural Intelligence. *CoRR* **2025**, *abs/2512.10624*, [arXiv:cs.CL/2512.10624]. https://doi.org/10.48550/arXiv.2512.10624.
74. Zhang, M.; Xu, Z.; Wang, P.; Li, R.; Wang, L.; Liu, Q.; Xu, J.; Zhang, X.; Wu, S.; Wang, L. AgriDoctor: A Multimodal Intelligent Assistant for Agriculture. *CoRR* **2025**, *abs/2509.17044*, [arXiv:cs.CV/2509.17044]. https://doi.org/10.48550/arXiv.2509.17044.
75. Ke, Y.; Yu, X.; Du, H.; Chapman, S.; Huang, H. Dynamic Orchestration of Multi-Agent System for Real-World Multi-Image Agricultural VQA. *CoRR* **2025**, *abs/2509.24350*, [arXiv:cs.CV/2509.24350]. https://doi.org/10.48550/arXiv.2509.24350.
76. Yang, B.; Zhang, Y.; Chen, Y.; Feng, L.; Xu, X.; Aierken, N.; Li, S. AgriAgent: Contract-Driven Planning and Capability-Aware Tool Orchestration in Real-World Agriculture. *CoRR* **2026**, *abs/2601.08308*, [arXiv:cs.CL/2601.08308]. https://doi.org/10.48550/arXiv.2601.08308.
77. Kandamali, D.F.; Porter, W.M.; Porter, E.; McLemore, A.; Rains, G.C. CottonBot: An AI-Driven Cotton Farming Assistant and Irrigation Advisor Using LLM-RAG and Agentic AI Tools. *Smart Agricultural Technology* **2025**, *12*, 101640. https://doi.org/10.1016/j.atech.2025.101640.
78. Geerts, M.; Reusens, M.; Baesens, B.; vanden Broucke, S.; Weerdt, J.D. On the Performance of LLMs for Real Estate Appraisal. *CoRR* **2025**, *abs/2506.11812*, [2506.11812].
79. Kvet, M.; Potočár, M.; Tatarka, S. Real Estate Attribute Value Extraction Using Large Language Models. *IEEE Access* **2025**, *13*, 13489–13506. https://doi.org/10.1109/ACCESS.2025.3564511.
80. Zhao, Y.; Gao, H. Utilizing Large Language Models for Information Extraction from Real Estate Transactions. *CoRR* **2024**, *abs/2404.18043*, [2404.18043]. https://doi.org/10.48550/ARXIV.2404.18043.
81. Wu, J.; Yang, C.; Wu, Y.; Mahns, S.; Wang, C.; Zhu, H.; Fang, F.; Xu, H. AI Realtor: Towards Grounded Persuasive Language Generation for Automated Copywriting, 2025, [arXiv:cs.CL/2502.16810].
82. Chung, J.; Ko, R.; Yoo, W.; Onizuka, M.; Kim, S.; Kim, T.W.; Shin, W.Y. GraphCompliance: Aligning Policy and Context Graphs for LLM-Based Regulatory Compliance. *arXiv preprint arXiv:2510.26309* **2025**.
83. Garza, L.; Elluri, L.; Kotal, A.; Piplai, A.; Gupta, D.; Joshi, A. Privcomp-kg: Leveraging knowledge graph and large language models for privacy policy compliance verification. *arXiv preprint arXiv:2404.19744* **2024**.
84. van Erven, O.; Zafeirakis, K.; Smit, J.; Smidi, J.; Buijs, L. [Re] Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. *Trans. Mach. Learn. Res.* **2025**, *2025*.
85. Park, J.S.; Zou, C.Q.; Shaw, A.; Hill, B.M.; Cai, C.J.; Morris, M.R.; Willer, R.; Liang, P.; Bernstein, M.S. Generative Agent Simulations of 1,000 People. *CoRR* **2024**, *abs/2411.10109*, [2411.10109]. https://doi.org/10.48550/ARXIV.2411.10109.
86. Pokharel, G.; Farabi, S.; Fowler, P.J.; Das, S. Street-Level AI: Are Large Language Models Ready for Real-World Judgments? In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025, Vol. 8, pp. 2043–2054.
87. Musumeci, E.; Brienza, M.; Suriani, V.; Nardi, D.; Bloisi, D.D. Llm based multi-agent generation of semi-structured documents from semantic templates in the public administration domain. In Proceedings of the International Conference on Human-Computer Interaction. Springer, 2024, pp. 98–117.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

88. de Castro Oliveira, J.V.; Rufino, N.; Levcovitz, S.; Utino, M.Y.R.; Lobato, F.; Turine, M.; Rezende, S.; Junior, A.F.L.J.; Marcacini, R.M. Agents4Gov: Privacy-Preserving Web Agents Using Open LLMs for Public Sector Tasks. In Proceedings of the Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). SBC, 2025, pp. 580–591.
89. Microsoft Corporation. Document Intelligence Invoice Model. Microsoft Learn Documentation, 2025. Documentation for Azure Document Intelligence's prebuilt invoice model, which uses OCR to analyze and extract structured invoice fields such as customer name, billing address, due date, amount due, and line items across various formats and languages.
90. Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. *arXiv (Cornell University)* **2020**. https://doi.org/10.48550/arxiv.2012.14740.
91. Clicteur, T.; Bertrand, D. Leveraging AI for Customs Classification Purposes. WCO News 102 – Issue 3 / 2023 (Point of View article), 2023. Article on how artificial intelligence can support Customs classification of products under the Harmonized System (HS) and improve trade compliance.
92. PricewaterhouseCoopers (PwC). Revolutionising Customs with AI. PDF Report, PricewaterhouseCoopers, 2024. Thought leadership report discussing how artificial intelligence can transform customs operations by enhancing operational efficiency, strengthening security, and enabling digital transformation across the customs value chain.
93. World Economic Forum. ChatWTO: An Analysis of Generative Artificial Intelligence and International Trade. White Paper, 2024. White paper exploring how generative artificial intelligence is transforming global trade, its opportunities, challenges and the need for international governance frameworks.
94. Mely.ai. AI-Powered Customs Automation Solution, 2025. Accessed: 2026-03-09.
95. Pazak, J.; Baeuml, M. Transform Supply Chain Logistics with Agentic AI. AWS Blog (Industries Category), 2025. AWS discusses how agentic AI and AI agents can be used to transform supply chain logistics operations by improving autonomy, responsiveness, and decision-making across inventory, warehousing, sourcing and other logistics workflows.
96. Aslett, J.; Cantens, T.; Chastel, F.; Crown, E.A.; Hamilton, S. Generative Artificial Intelligence for Compliance Risk Analysis: Applications in Tax and Customs Administration. Technical notes and manuals no. 2025/013, International Monetary Fund, 2025. Technical guidance on using generative artificial intelligence (GenAI) for compliance risk analysis in tax and customs administration, https://doi.org/10.5089/9798229012430.005.
97. Anthropic. Project Vend: Can Claude Run a Small Shop? (And Why Does That Matter?). https://www.anthropic.com/research/project-vend-1, 2025. Accessed: 2026-04-25.
98. Lin, P.; Tsai, Y. ScoreRAG: A Retrieval-Augmented Generation Framework with Consistency-Relevance Scoring and Structured Summarization for News Generation. *CoRR* **2025**, *abs/2506.03704*, [2506.03704]. https://doi.org/10.48550/ARXIV.2506.03704.
99. Dharmavaram, A.; Hakak, S. SANCTUARY: An Efficient Evidence-based Automated Fact Checking System. In Proceedings of the Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER); Akhtar, M.; Aly, R.; Christodoulopoulos, C.; Cocarascu, O.; Guo, Z.; Mittal, A.; Schlichtkrull, M.; Thorne, J.; Vlachos, A., Eds., Vienna, Austria, 2025; pp. 247–257. https://doi.org/10.18653/v1/2025.fever-1.19.
100. Lee, D.; Choi, J.; Han, J.; Park, K. Journalism-Guided Agentic In-Context Learning for News Stance Detection. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 15404–15427.
101. Cui, Z.; Huang, T.; Chiang, C.E.; Du, C. Toward verifiable misinformation detection: A multi-tool LLM agent framework. In Proceedings of the Proceedings of the 2025 International Conference on Generative Artificial Intelligence for Business, 2025, pp. 179–185.
102. Iqbal, H.; Wang, Y.; Wang, M.; Georgiev, G.N.; Geng, J.; Gurevych, I.; Nakov, P. OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - System Demonstrations, Miami, Florida, USA, November 12-16, 2024; Farías, D.I.H.; Hope, T.; Li, M., Eds. Association for Computational Linguistics, 2024, pp. 219–229. https://doi.org/10.18653/V1/2024.EMNLP-DEMO.23.
103. Sandoval-Castaneda, M.; Russell, B.; Sivic, J.; Shakhnarovich, G.; Caba Heilbron, F. EditDuet: A Multi-Agent System for Video Non-Linear Editing. In Proceedings of the Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers, 2025, pp. 1–11.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

104. Jiang, G.; Shi, X.; Luo, Q. JRE-L: Journalist, Reader, and Editor LLMs in the Loop for Science Journalism for the General Audience. In Proceedings of the North American Chapter of the Association for Computational Linguistics, 2025.
105. Ouyang, L. Can Memory-Augmented LLM Agents Aid Journalism in Interpreting and Framing News for Diverse Audiences?, 2025, [arXiv:cs.CY/2507.21055].
106. Wang, Z.; Yu, Y.; Zheng, W.; Ma, W.; Zhang, M. Macrec: A multi-agent collaboration framework for recommendation. In Proceedings of the Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2760–2764.
107. Zhang, A.; Chen, Y.; Sheng, L.; Wang, X.; Chua, T.S. On generative agents in recommendation. In Proceedings of the Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval, 2024, pp. 1807–1817.
108. Xu, Z.; Wang, L.; Wang, J.; Li, Z.; Shi, S.; Yang, X.; Wang, Y.; Hu, B.; Yu, J.; Zhang, M. Filmagent: A multi-agent framework for end-to-end film automation in virtual 3d spaces. *arXiv preprint arXiv:2501.12909* **2025**.
109. He, L.; Song, Y.; Huang, H.; Liu, P.; Tang, Y.; Aliaga, D.; Zhou, X. Kubrick: Multimodal agent collaborations for synthetic video generation. *arXiv preprint arXiv:2408.10453* **2024**.
110. Wu, W.; Zhu, Z.; Shou, M.Z. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314* **2025**.
111. Ni, M.; Wu, C.; Yuan, H.; Yang, Z.; Gong, M.; Wang, L.; Liu, Z.; Zuo, W.; Duan, N. AutoDirector: Online Auto-scheduling Agents for Multi-sensory Composition. *arXiv preprint arXiv:2408.11564* **2024**.
112. Swacha, J.; Gracel, M. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Applied Sciences* **2025**, *15*, 4234. https://doi.org/10.3390/app15084234.
113. Dan, Y.; Lei, Z.; Gu, Y.; Li, Y.; Yin, J.; Lin, J.; Ye, L.; Tie, Z.; Zhou, Y.; Wang, Y.; et al. EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. *arXiv* **2023**, [arXiv:cs.CL/2308.02773].
114. LearnLM Team, Google. LearnLM: Improving Gemini for Learning. *arXiv preprint arXiv:2412.16429* **2024**. https://doi.org/10.48550/arXiv.2412.16429.
115. Lee, U.; Lee, J.; Bae, J.; Jeong, Y.; Koh, J.; Lee, G.B.; Lee, G.; Ahn, T.; Kim, H. Pedagogy-R1: Pedagogical Large Reasoning Model and Well-balanced Educational Benchmark. In Proceedings of the Proceedings of the 34th ACM International Conference on Information and Knowledge Management, Seoul, Republic of Korea, 2025; CIKM '25, pp. 1497–1507. https://doi.org/10.1145/3746252.3761133.
116. Chu, Y.; Li, H.; Yang, K.; Copur-Gencturk, Y.; Tang, J. LLM-based Automated Grading with Human-in-the-Loop. *arXiv preprint arXiv:2504.05239* **2025**. Accepted to IEEE TALE 2025, https://doi.org/10.48550/arXiv.2504.05239.
117. Fajardo-Ramos, D.C.; Chiappe, A.; Mella-Norambuena, J. Human-in-the-loop assessment with AI: implications for teacher education in Ibero-American universities. In Proceedings of the Frontiers in Education. Frontiers Media SA, 2025, Vol. 10, p. 1710992.
118. Liu, J.; Huang, Z.; Xiao, T.; Sha, J.; Wu, J.; Liu, Q.; Wang, S.; Chen, E. SocraticLM: exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems* **2024**, *37*, 85693–85721.
119. Gao, W.; Liu, Q.; Yue, L.; Yao, F.; Lv, R.; Zhang, Z.; Wang, H.; Huang, Z. Agent4Edu: Generating Learner Response Data by Generative Agents for Intelligent Education Systems. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 23923–23932. https://doi.org/10.1609/aaai.v39i22.34565.
120. Xu, S.; Wen, H.N.; Pan, H.; Dominguez, D.; Hu, D.; Zhang, X. Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. In Proceedings of the Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–26.
121. Guo, S.; Latif, E.; Zhou, Y.; Huang, X.; Zhai, X. Using Generative AI and Multi-Agents to Provide Automatic Feedback. *arXiv preprint arXiv:2411.07407* **2024**. https://doi.org/10.48550/arXiv.2411.07407.
122. Macina, J.; Daheim, N.; Hakimi, I.; Kapur, M.; Gurevych, I.; Sachan, M. MathTutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 2025; pp. 204–221. https://doi.org/10.18653/v1/2025.emnlp-main.11.
123. Chu, Z.; Wang, S.; Xie, J.; Zhu, T.; Yan, Y.; Ye, J.; Zhong, A.; Hu, X.; Liang, J.; Yu, P.S.; et al. LLM Agents for Education: Advances and Applications. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 2025; pp. 13782–13810.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

124. Nedungadi, P.; Tang, K.Y.; Raman, R. The Transformative Power of Generative Artificial Intelligence for Achieving the Sustainable Development Goal of Quality Education. *Sustainability* **2024**, *16*, 9779. https://doi.org/10.3390/su16229779.
125. Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; Yu, P.S.; Wen, Q. Large Language Models for Education: A Survey and Outlook. *arXiv preprint arXiv:2403.18105* **2024**. https://doi.org/10.48550/arXiv.2403.18105.
126. Delikoura, I.; Fung, Y.R.M.; Hui, P. From Superficial Outputs to Superficial Learning: Risks of Large Language Models in Education. *arXiv preprint arXiv:2509.21972* **2025**. https://doi.org/10.48550/arXiv.2509.21972.
127. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
128. Zhou, Z.; Shi, J.; Song, P.; Yang, X.; Jin, Y.; Guo, L.; Li, Y. LawGPT: A Chinese Legal Knowledge-Enhanced Large Language Model. *CoRR* **2024**, *abs/2406.04614*, [2406.04614]. https://doi.org/10.48550/ARXIV.2406.04614.
129. Huang, Q.; Tao, M.; An, Z.; Zhang, C.; Jiang, C.; Chen, Z.; Wu, Z.; Feng, Y. Lawyer LLaMA Technical Report. *CoRR* **2023**, *abs/2305.15062*, [2305.15062]. https://doi.org/10.48550/ARXIV.2305.15062.
130. Cui, J.; Li, Z.; Yan, Y.; Chen, B.; Yuan, L. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. *CoRR* **2023**, *abs/2306.16092*, [2306.16092]. https://doi.org/10.48550/ARXIV.2306.16092.
131. Colombo, P.; Pires, T.P.; Boudiaf, M.; Culver, D.; Melo, R.; Corro, C.; Martins, A.F.T.; Esposito, F.; Raposo, V.L.; Morgado, S.; et al. SaulLM-7B: A pioneering Large Language Model for Law. *CoRR* **2024**, *abs/2403.03883*, [2403.03883]. https://doi.org/10.48550/ARXIV.2403.03883.
132. Fei, Z.; Zhang, S.; Shen, X.; Zhu, D.; Wang, X.; Ge, J.; Ng, V. InternLM-Law: An Open-Sourced Chinese Legal Large Language Model. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025; Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B.D.; Schockaert, S., Eds. Association for Computational Linguistics, 2025, pp. 9376–9392.
133. Rawte, V.; Sheth, A.P.; Das, A. A Survey of Hallucination in Large Foundation Models. *CoRR* **2023**, *abs/2309.05922*, [2309.05922]. https://doi.org/10.48550/ARXIV.2309.05922.
134. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual; Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; Lin, H., Eds., 2020.
135. Shi, J.; Guo, Q.; Liao, Y.; Wang, Y.; Chen, S.; Liang, S. Legal-LM: Knowledge Graph Enhanced Large Language Models for Law Consulting. In Proceedings of the Advanced Intelligent Computing Technology and Applications - 20th International Conference, ICIC 2024, Tianjin, China, August 5-8, 2024, Proceedings, Part IV (LNAI); Huang, D.; Si, Z.; Zhang, C., Eds. Springer, 2024, Vol. 14878, *Lecture Notes in Computer Science*, pp. 175–186. https://doi.org/10.1007/978-981-97-5672-8_15.
136. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
137. Yue, S.; Chen, W.; Wang, S.; Li, B.; Shen, C.; Liu, S.; Zhou, Y.; Xiao, Y.; Yun, S.; Huang, X.; et al. DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services. *CoRR* **2023**, *abs/2309.11325*, [2309.11325]. https://doi.org/10.48550/ARXIV.2309.11325.
138. Wang, Z.; Yuan, B. L-MARS: Legal Multi-Agent Workflow with Orchestrated Reasoning and Agentic Search. *CoRR* **2025**, *abs/2509.00761*, [2509.00761]. https://doi.org/10.48550/ARXIV.2509.00761.
139. Raptopoulos, P.; Filandrianos, G.; Lymperaiou, M.; Stamou, G. PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds. Association for Computational Linguistics, 2025, pp. 7948–7984. https://doi.org/10.18653/V1/2025.EMNLP-MAIN.403.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

140. Watson, W.; Cho, N.; Srishankar, N.; Zeng, Z.; Cecchi, L.; Scott, D.; Siddagangappa, S.; Kaur, R.; Balch, T.; Veloso, M. LAW: Legal Agentic Workflows for Custody and Fund Services Contracts. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025 - Industry Track, Abu Dhabi, UAE, January 19-24, 2025; Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B.D.; Schockaert, S.; Darwish, K.; Agarwal, A., Eds. Association for Computational Linguistics, 2025, pp. 583–594.
141. He, Z.; Cao, P.; Wang, C.; Jin, Z.; Chen, Y.; Xu, J.; Li, H.; Liu, K.; Zhao, J. AgentsCourt: Building Judicial Decision-Making Agents with Court Debate Simulation and Legal Knowledge Augmentation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y., Eds. Association for Computational Linguistics, 2024, Vol. EMNLP 2024, *Findings of ACL*, pp. 9399–9416. https://doi.org/10.18653/V1/2024.FINDINGS-EMNLP.549.
142. Wang, Y.; Chen, Y.; Meng, F.; Chen, X.; Yang, X.; Zhang, M. Law in Silico: Simulating Legal Society with LLM-Based Agents. *CoRR* **2025**, *abs/2510.24442*, [2510.24442]. https://doi.org/10.48550/ARXIV.2510.24442.
143. Luo, Z.; Xu, C.; Zhao, P.; Sun, Q.; Geng, X.; Hu, W.; Tao, C.; Ma, J.; Lin, Q.; Jiang, D. WizardCoder: Empowering Code Large Language Models with Evol-Instruct. In Proceedings of the The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024.
144. Le, H.; Wang, Y.; Gotmare, A.D.; Savarese, S.; Hoi, S.C. CodeRL: Mastering Code Generation through Pretrained Models and Deep Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
145. Shojaee, P.; Jain, A.; Tipirneni, S.; Reddy, C.K. Execution-based Code Generation using Deep Reinforcement Learning. *Trans. Mach. Learn. Res.* **2023**, *2023*.
146. Dou, S.; Liu, Y.; Jia, H.; Xiong, L.; Zhou, E.; Shen, W.; Shan, J.; Huang, C.; Wang, X.; Fan, X.; et al. StepCoder: Improve Code Generation with Reinforcement Learning from Compiler Feedback. *CoRR* **2024**, *abs/2402.01391*, [2402.01391]. https://doi.org/10.48550/ARXIV.2402.01391.
147. DeepSeek-AI.; Zhu, Q.; Guo, D.; Shao, Z.; Yang, D.; Wang, P.; Xu, R.; Wu, Y.; Li, Y.; Gao, H.; et al. DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence. *CoRR* **2024**, *abs/2406.11931*, [2406.11931]. https://doi.org/10.48550/ARXIV.2406.11931.
148. Chen, X.; Lin, M.; Schärli, N.; Zhou, D. Teaching Large Language Models to Self-Debug. In Proceedings of the The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024.
149. Chen, X.; Tao, Z.; Zhang, K.; Zhou, C.; Zhang, X.; Gu, W.; He, Y.; Zhang, M.; Cai, X.; Zhao, H.; et al. Revisit Self-Debugging with Self-Generated Tests for Code Generation. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds. Association for Computational Linguistics, 2025, pp. 18003–18023.
150. Adnan, M.; Xu, Z.; Kuhn, C.C.N. Large Language Model Guided Self-Debugging Code Generation. *CoRR* **2025**, *abs/2502.02928*, [2502.02928]. https://doi.org/10.48550/ARXIV.2502.02928.
151. Wen, J.; Guan, J.; Wang, H.; Wu, W.; Huang, M. CodePlan: Unlocking Reasoning Potential in Large Language Models by Scaling Code-form Planning. In Proceedings of the The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net, 2025.
152. Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; et al. ChatDev: Communicative Agents for Software Development. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024; Ku, L.; Martins, A.; Srikumar, V., Eds. Association for Computational Linguistics, 2024, pp. 15174–15186. https://doi.org/10.18653/V1/2024.ACL-LONG.810.
153. Islam, M.A.; Ali, M.E.; Parvez, M.R. MapCoder: Multi-Agent Code Generation for Competitive Problem Solving. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024; Ku, L.; Martins, A.; Srikumar, V., Eds. Association for Computational Linguistics, 2024, pp. 4912–4944. https://doi.org/10.18653/V1/2024.ACL-LONG.269.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

154. Pan, J.; Wang, X.; Neubig, G.; Jaitly, N.; Ji, H.; Suhr, A.; Zhang, Y. Training Software Engineering Agents and Verifiers with SWE-Gym. In Proceedings of the Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025; Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; Zhu, J., Eds. PMLR / OpenReview.net, 2025, Vol. 267, *Proceedings of Machine Learning Research*.
155. Hu, M.; Zhao, P.; Xu, C.; Sun, Q.; Lou, J.G.; Lin, Q.; Luo, P.; Rajmohan, S. Agentgen: Enhancing planning abilities for large language model based agent via environment and task generation. In Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, 2025, pp. 496–507.
156. Jain, N.; Singh, J.; Shetty, M.; Zheng, L.; Sen, K.; Stoica, I. R2E-Gym: Procedural Environments and Hybrid Verifiers for Scaling Open-Weights SWE Agents. *CoRR* **2025**, *abs/2504.07164*, [2504.07164]. https://doi.org/10.48550/ARXIV.2504.07164.
157. Luo, M.; Jain, N.; Singh, J.; Tan, S.; Patel, A.; Wu, Q.; Ariyak, A.; Cai, C.; Tarun Venkat, S.Z.; Athiwaratkun, B.; et al. DeepSWE: Training a State-of-the-Art Coding Agent from Scratch by Scaling RL. https://pretty-radio-b75.notion.site/DeepSWE-Training-a-Fully-Open-sourced-State-of-the-Art-Coding-Agent-by-Scaling-RL-22281902c1468193aabbe9a8c59bbe33, 2025. Notion Blog.
158. Chen, Z.; Zhao, C.; Chen, B.; Lin, D.; Chen, Y.; Leung, A.; Rajbahadur, G.K.; Oliva, G.A.; Zhang, H.; Bhatia, A.; et al. RepoForge: Training a SOTA Fast-thinking SWE Agent with an End-to-End Data Curation Pipeline Synergizing SFT and RL at Scale. *CoRR* **2025**, *abs/2508.01550*, [2508.01550]. https://doi.org/10.48550/ARXIV.2508.01550.
159. Lee, J.; Stevens, N.; et al. Large Language Models in Finance: A Survey. *Neural Computing and Applications* **2024**. https://doi.org/10.1007/s00521-024-10495-6.
160. Sanz-Guerrero, M.; Arroyo, J. Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in P2P Lending. *arXiv preprint arXiv:2401.16458* **2024**, [2401.16458]. https://doi.org/10.48550/arXiv.2401.16458.
161. Liu, Z.; Guo, X.; Yang, Z.; Lou, F.; Zeng, L.; Niu, J.; Li, M.; Qi, Q.; Liu, Z.; Han, Y.; et al. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning. *arXiv preprint arXiv:2503.16252* **2025**, [arXiv:cs.CL/2503.16252]. https://doi.org/10.48550/arXiv.2503.16252.
162. Zhu, J.; Chen, Q.; Dou, H.; Li, J.; Guo, L.; Chen, F.; Zhang, C. DianJin-R1: Evaluating and Enhancing Financial Reasoning in Large Language Models. *arXiv preprint arXiv:2504.15716* **2025**, [arXiv:cs.AI/2504.15716]. https://doi.org/10.48550/arXiv.2504.15716.
163. Geng, W. Could Large Language Models work as Post-hoc Explainability tools in Credit Risk Model Governance? *arXiv preprint arXiv:2602.18895* **2026**, [2602.18895]. https://doi.org/10.48550/arXiv.2602.18895.
164. Dong, Y.; Wu, F.; Zhang, K.; Dai, Y.; Zhang, S.; Ye, W.; Chen, S.; Cheng, Z.Q. Large Language Model Agents in Finance: A Survey Bridging Research, Practice, and Real-World Deployment. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 2025; pp. 17889–17907. https://doi.org/10.18653/v1/2025.findings-emnlp.972.
165. Evite, P.M.; Svetlova, E.; Bucur, D. Trade-offs in Financial AI: Explainability in a Trilemma with Accuracy and Compliance. *arXiv preprint arXiv:2602.01368* **2026**, [2602.01368]. https://doi.org/10.48550/arXiv.2602.01368.
166. Yang, H.; Zhang, B.; Wang, N.; Guo, C.; Zhang, X.; Lin, L.; Wang, J.; Zhou, T.; Guan, M.; Zhang, R.; et al. FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models. *arXiv preprint arXiv:2405.14767* **2024**, [arXiv:q-fin.ST/2405.14767]. https://doi.org/10.48550/arXiv.2405.14767.
167. Cai, T.; Li, G.; Han, N.; Huang, C.; Wang, Z.; Zeng, C.; Wang, Y.; Zhou, J.; Zhang, H.; Chen, Q.; et al. FinDebate: Multi-Agent Collaborative Intelligence for Financial Analysis. *arXiv preprint arXiv:2509.17395* **2025**, [arXiv:cs.CL/2509.17395]. https://doi.org/10.48550/arXiv.2509.17395.
168. Jin, J.; Zhang, Y.; Xu, Y.; Qian, H.; Zhu, Y.; Dou, Z. FinSight: Towards Real-World Financial Deep Research. *arXiv preprint arXiv:2510.16844* **2025**, [2510.16844]. https://doi.org/10.48550/arXiv.2510.16844.
169. Zhang, W.; Zhao, L.; Xia, H.; Sun, S.; Sun, J.; Qin, M.; Li, X.; Zhao, Y.; Zhao, Y.; Cai, X.; et al. A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist. *arXiv preprint arXiv:2402.18485* **2024**, [arXiv:q-fin.TR/2402.18485]. https://doi.org/10.48550/arXiv.2402.18485.
170. Xiao, Y.; Sun, E.; Luo, D.; Wang, W. TradingAgents: Multi-Agents LLM Financial Trading Framework. *arXiv preprint arXiv:2412.20138* **2024**, [2412.20138]. https://doi.org/10.48550/arXiv.2412.20138.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

171. Zhao, L.; Sun, R.; Jiang, Z.; Yang, B.; Bai, Y.; Chen, M.; Wang, X.; Li, J.; Bai, Z. ContestTrade: A Multi-Agent Trading System Based on Internal Contest Mechanism. *arXiv preprint arXiv:2508.00554* **2025**, [arXiv:q-fin.TR/2508.00554]. https://doi.org/10.48550/arXiv.2508.00554.
172. Xiong, F.; Zhang, X.; Feng, A.; Sun, S.; You, C. QuantAgent: Price-Driven Multi-Agent LLMs for High-Frequency Trading. *arXiv preprint arXiv:2509.09995* **2025**, [arXiv:cs.CE/2509.09995]. https://doi.org/10.48550/arXiv.2509.09995.
173. Suryavanshi, K.P.; kumar ks, S. AI-Driven Financial Risk Management with LLM Agents and Adaptive Learning. ResearchGate Preprint, 2025. https://doi.org/10.13140/RG.2.2.19050.22728/1.
174. Cao, Y.; Chen, Z.; Pei, Q.; Dimino, F.; Ausiello, L.; Kumar, P.; Subbalakshmi, K.P.; Ndiaye, P.M. RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data. *arXiv preprint arXiv:2404.07452* **2024**, [2404.07452]. https://doi.org/10.48550/arXiv.2404.07452.
175. McKinsey & Company. The End of Inertia: Agentic AI's Disruption of Retail and SME Banking. McKinsey article / industry report, 2025.
176. Miao, J.; Su, Z.; Xiao, X. MMLM-CA: A Multimodal Large Language Model Framework for Automated Insurance Claims Adjudication Integrating Policy Document Understanding and Visual Damage Assessment. In Proceedings of the Proceedings of the 2025 3rd International Conference on Artificial Intelligence, Systems and Network Security (AISNS 2025), 2026. https://doi.org/10.1145/3797161.3797170.
177. Asgarian, A.; Saha, R.; Jakubovitz, D.; Peyre, J. AutoFraudNet: A Multimodal Network to Detect Fraud in the Auto Insurance Industry. *arXiv preprint arXiv:2301.07526* **2023**, [2301.07526]. https://doi.org/10.48550/arXiv.2301.07526.
178. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning* **2021**, *14*, 1–210. https://doi.org/10.1561/2200000083.
179. Wu, C.; Lin, W.; Zhang, X.; Zhang, Y.; Xie, W.; Wang, Y. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association* **2024**, *31*, 1833–1843. https://doi.org/10.1093/jamia/ocae045.
180. Lin, X.; Xu, C.; Xiong, Z.; Zhang, X.; Ni, N.; Ni, B.; Chang, J.; Pan, R.; Wang, Z.; Yu, F.; et al. PanGu Drug Model: learn a molecule like a human. *Science China Life Sciences* **2022**, *66*, 879–882. https://doi.org/10.1007/s11427-022-2239-y.
181. Wang, X.; Li, J.; Chen, S.; Zhu, Y.; Wu, X.; Zhang, Z.; Xu, X.; Chen, J.; Fu, J.; Wan, X.; et al. Huatuo-26M, a large-scale Chinese medical QA dataset. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 3828–3848.
182. Wang, H.; Zhao, S.; Qiang, Z.; Li, Z.; Liu, C.; Xi, N.; Du, Y.; Qin, B.; Liu, T. Knowledge-tuning Large Language Models with Structured Medical Knowledge Bases for Trustworthy Response Generation in Chinese. *ACM Trans. Knowl. Discov. Data* **2025**. Just Accepted, https://doi.org/10.1145/3686807.
183. Lan, W.; Wang, W.; Ji, C.; Yang, G.; Zhang, Y.; Liu, X.; Wu, S.; Wang, G. ClinicalGPT-R1: Pushing reasoning capability of generalist disease diagnosis with large language model. *arXiv (Cornell University)* **2025**. https://doi.org/https://doi.org/10.48550/arxiv.2504.09421.
184. Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; Wang, B. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. *arXiv (Cornell University)* **2024**, *abs/2412.18925*. https://doi.org/10.48550/arxiv.2412.18925.
185. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* **2022**.
186. Yossi Matias, Aashima Gupta. Introducing MedLM for the healthcare industry. Google Cloud Blog (Healthcare & Life Sciences), 2023. Blog announcing MedLM, a family of foundation models fine-tuned for healthcare use cases, now available to Google Cloud customers through Vertex AI platform.
187. Kees Hertogh. Meet Microsoft Dragon Copilot: Your New AI Assistant for Clinical Workflow. Microsoft Industry Blog, 2025. Announcement of Microsoft Dragon Copilot, an AI-powered assistant designed to streamline clinical workflows by automating documentation, surfacing key information, and reducing administrative burden for clinicians.
188. Microsoft. Dragon Copilot: AI Assistant for Clinical Workflow. https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-copilot, 2025. Microsoft Cloud for Healthcare solution page. Accessed: 2026-03-09.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

189. Amazon Web Services. Introducing AWS HealthScribe – automatically generate clinical notes from patient-clinician conversations using AWS HealthScribe. https://aws.amazon.com/blogs/industries/industries-introducing-aws-healthscribe/, 2023. AWS Industries Blog, Accessed: 2026-03-09.
190. Li, J.; Wang, S.; Zhang, M.; Li, W.; Lai, Y.; Kang, X.; Ma, W.; Liu, Y. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv (Cornell University)* **2024**, *abs/2405.02957*. https://doi.org/10.48550/arxiv.2405.02957.
191. Cho, A.; Woo, J.; Shi, B.; Udeshi, A.; Woo, J. The Application of MATEC (Multi-AI Agent Team Care) Framework in Sepsis Care. *arXiv (Cornell University)* **2025**, *abs/2503.16433*. https://doi.org/10.48550/arxiv.2503.16433.
192. Wang, Z.; Wu, J.; Low, C.; Cai, L.; Yang, X.; Li, Q.; Jin, Y. MedAgent-Pro: Towards Multi-modal Evidence-based Medical Diagnosis via Reasoning Agentic Workflow. *arXiv (Cornell University)* **2025**, *abs/2503.18968*. https://doi.org/10.48550/arxiv.2503.18968.
193. Vichentijevikj, I.; Mishev, K.; Misheva, M. Prompt-to-Pill: Multi-Agent Drug Discovery and Clinical Simulation Pipeline. *Bioinformatics Advances* **2025**, *6*, vbaf323–vbaf323. https://doi.org/10.1093/bioadv/vbaf323.
194. Zhang, H.G.; Eckmann, P.; Miao, J.; Mahon, A.B.; Zou, J. The Virtual Biotech: A Multi-Agent AI Framework for Therapeutic Discovery and Development. *bioRxiv* **2026**. bioRxiv preprint.
195. OpenAI. AI as a Healthcare Ally: How Americans Are Navigating the System with ChatGPT. Technical report, OpenAI, 2026. Accessed: 2026-03-04.
196. Wang, S.; Yuan, H.; Zhou, L.; Ni, L.M.; Shum, H.Y.; Guo, J. Alpha-GPT: Human-AI Interactive Alpha Mining for Quantitative Investment. *arXiv preprint arXiv:2308.00016* **2023**, [2308.00016]. https://doi.org/10.48550/arXiv.2308.00016.
197. Wen, Y.L.; Chen, X. X-GridAgent: An LLM-Powered Agentic AI System for Assisting Power Grid Analysis. *arXiv preprint arXiv:2512.20789* **2025**, [arXiv:eess.SY/2512.20789]. https://doi.org/10.48550/arXiv.2512.20789.
198. Jin, H.; Kim, K.; Kwon, J. GridMind: LLMs-Powered Agents for Power System Analysis and Operations. In Proceedings of the SC Workshops '25: Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, St. Louis, MO, USA, 2025; pp. 560–568. https://doi.org/10.1145/3731599.3767409.
199. Bernier, F.; Cao, J.; Cordy, M.; Ghamizi, S. PowerGraph-LLM: Novel Power Grid Graph Embedding and Optimization With Large Language Models. *IEEE Transactions on Power Systems* **2025**, *40*, 5483–5486. https://doi.org/10.1109/TPWRS.2025.3596774.
200. Zhang, Y.; Saber, A.M.; Youssef, A.; Kundur, D. Grid-Agent: An LLM-Powered Multi-Agent System for Power Grid Control. *arXiv preprint arXiv:2508.05702* **2025**, [arXiv:cs.MA/2508.05702]. https://doi.org/10.48550/arXiv.2508.05702.
201. Cheng, Y.; Zhao, H.; Zhou, X.; Zhao, J.; Cao, Y.; Yang, C. GAIA – A Large Language Model for Advanced Power Dispatch. *arXiv preprint arXiv:2408.03847* **2024**, [arXiv:eess.SY/2408.03847]. https://doi.org/10.48550/arXiv.2408.03847.
202. Christou, P.; Islam, M.Z.; Lin, Y.; Xiong, J. LLM4DistReconfig: A Fine-tuned Large Language Model for Power Distribution Network Reconfiguration. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Albuquerque, New Mexico, 2025; pp. 4136–4155. https://doi.org/10.18653/v1/2025.naacl-long.208.
203. Jeong, M.; Chang, J.; Yoon, Y. Speak to Simulate: An LLM-Guided Agentic Framework for Traffic Simulation in SUMO. In Proceedings of the Proceedings of the 8th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation (GeoSIM '25). Association for Computing Machinery, 2025, pp. 45–48. https://doi.org/10.1145/3764921.3770151.
204. Du, Y.; Zhang, J.; Feng, J.; Liu, Z.; Yuan, J.; Li, Y. TrafficSimAgent: A Hierarchical Agent Framework for Autonomous Traffic Simulation with MCP Control. *arXiv preprint arXiv:2512.20996* **2025**, [arXiv:cs.AI/2512.20996]. https://doi.org/10.48550/arXiv.2512.20996.
205. Lai, S.; Xu, Z.; Zhang, W.; Liu, H.; Xiong, H. LLMLight: Large Language Models as Traffic Signal Control Agents. In Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25). Association for Computing Machinery, 2025, pp. 2335–2346. https://doi.org/10.1145/3690624.3709379.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

206. Hottung, A.; Berto, F.; Hua, C.; Gast Zepeda, N.; Wetzel, D.; Römer, M.; Ye, H.; Zago, D.; Poli, M.; Massaroli, S.; et al. VRPAgent: LLM-Driven Discovery of Heuristic Operators for Vehicle Routing Problems. *arXiv preprint arXiv:2510.07073* **2025**, [arXiv:cs.AI/2510.07073]. https://doi.org/10.48550/arXiv.2510.07073.
207. Zhang, N.; Cao, Z.; Zhou, J.; Zhang, C.; Ong, Y.S. An Agentic Framework with LLMs for Solving Complex Vehicle Routing Problems. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026. Poster.
208. Aransyah, M.F.; Hermanto, B.; Muftiadi, A.; Oktadiana, H. Conversational AI in hospitality and tourism: a bibliometric–systematic review. *Cogent Business & Management* **2026**, *13*, 2613599. https://doi.org/10.1080/23311975.2026.2613599.
209. Talón-Ballestero, P.; Nieto-García, M.; González-Serrano, L. The wheel of dynamic pricing: towards open pricing and one to one pricing in hotel revenue management. *International Journal of Hospitality Management* **2022**, *102*, 103184. https://doi.org/10.1016/j.ijhm.2022.103184.
210. Henriques, H.; Pereira, L.N. Hotel demand forecasting models and methods using artificial intelligence: A systematic literature review. *Tourism & Management Studies* **2024**, *20*, 39–51. https://doi.org/10.18089/tms.20240304.
211. Noh, D.; Nam, H.; Gillespie, K.; Liu, Y.; Hong, D. Yummy Operations Robot Initiative: An Autonomous Cooking System Utilizing a Modular Robotic Kitchen and Dual-Arm Proprioceptive Manipulator. *IEEE Robotics & Automation Magazine* **2025**, pp. 2–14. https://doi.org/10.1109/MRA.2025.3615353.
212. Sakib, M.S.; Sun, Y. From Cooking Recipes to Robot Task Trees – Improving Planning Correctness and Task Efficiency by Leveraging LLMs with a Knowledge Network. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 12704–12711. https://doi.org/10.1109/ICRA57147.2024.10611369.
213. Kanazawa, N.; Kawaharazuka, K.; Obinata, Y.; Okada, K.; Inaba, M. Real-world cooking robot system from recipes based on food state recognition using foundation models and PDDL. *Advanced Robotics* **2024**, *38*, 1318–1334. https://doi.org/10.1080/01691864.2024.2407136.
214. Wang, Q.; Huang, Z.; Jia, R.; Debevec, P.; Yu, N. MAViS: A Multi-Agent Framework for Long-Sequence Video Storytelling. In Proceedings of the Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Rabat, Morocco, 2026; pp. 2273–2295. https://doi.org/10.18653/v1/2026.eacl-long.101.
215. Tsiavos, V.; Kitsios, F.C. The digital transformation of the film industry: How Artificial Intelligence is changing the seventh art. *Telecommunications Policy* **2025**, *49*, 103021. https://doi.org/10.1016/j.telpol.2025.103021.
216. Wu, Z.; Chen, Z.; Zhu, D.; Mousas, C.; Kao, D. A Systematic Review of Generative AI on Game Character Creation: Applications, Challenges, and Future Trends. *IEEE Transactions on Games* **2025**, pp. 1–15. Early Access / in press, https://doi.org/10.1109/TG.2025.3564869.
217. Christiansen, F.R.; Hollensberg, L.N.; Jensen, N.B.F.; Julsgaard, K.; Jespersen, K.N.; Nikolov, I.A. Exploring Presence in Interactions with LLM-Driven NPCs: A Comparative Study of Speech Recognition and Dialogue Options. In Proceedings of the Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology, 2024, pp. 1–11. https://doi.org/10.1145/3641825.3687716.
218. Kim, B.; Kim, M.; Seo, D.; Kim, B. Leveraging Large Language Models for Active Merchant Non-player Characters. In Proceedings of the Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 10108–10116. https://doi.org/10.24963/ijcai.2025/1123.
219. Wen, J.; Ma, B. Enhancing museum experience through deep learning and multimedia technology. *Heliyon* **2024**, *10*, e32706. https://doi.org/10.1016/j.heliyon.2024.e32706.
220. Al-Busaidi, A.S.; Raman, R.; Hughes, L.; Albashrawi, M.A.; Malik, T.; Dwivedi, Y.K.; Al-Alawi, T.; AlRizeiqi, M.; Davies, G.; Fenwick, M.; et al. Redefining boundaries in innovation and knowledge domains: Investigating the impact of generative artificial intelligence on copyright and intellectual property rights. *Journal of Innovation & Knowledge* **2024**, *9*, 100630. https://doi.org/10.1016/j.jik.2024.100630.
221. Amankwah-Amoah, J.; Abdalla, S.; Mogaji, E.; Elbanna, A.; Dwivedi, Y.K. The impending disruption of creative industries by generative AI: Opportunities, challenges, and research agenda. *International Journal of Information Management* **2024**, *79*, 102759. https://doi.org/10.1016/j.ijinfomgt.2024.102759.
222. Cyberhaven Labs. 2025 AI Adoption and Risk Report. Technical report, Cyberhaven, 2025. Accessed: 2026-04-12.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

223. Menlo Security. Menlo Security's 2025 Report Uncovers 68% Surge in "Shadow" Generative AI Usage in the Modern Enterprise. Technical report, Menlo Security, 2025. Accessed: 2026-04-12.
224. A10 Networks. Generative AI Security Risks, 2025. Accessed: 2026-04-12.
225. Practical DevSecOps. AI Security Statistics 2026: Latest Data, Trends & Research, 2026. Accessed: 2026-04-12.
226. Wang, W.; Han, D.; Diaz, D.; Xu, J.; Rühle, V.; Rajmohan, S. OdysseyBench: Evaluating LLM Agents on Long-Horizon Complex Office Application Workflows. *arXiv (Cornell University)* **2025**, *abs/2508.09124*. https://doi.org/10.48550/arxiv.2508.09124.
227. Abhyankar, R.; Qi, Q.; Zhang, Y. OSWorld-Human: Benchmarking the Efficiency of Computer-Use Agents. *arXiv (Cornell University)* **2025**, *abs/2506.16042*. https://doi.org/10.48550/arxiv.2506.16042.
228. Atil, B.; Chittams, A.; Fu, L.; Ture, F.; Xu, L.; Baldwin, B. Non-Determinism of "Deterministic" LLM Settings. *arXiv (Cornell University)* **2024**, *abs/2408.04667*. https://doi.org/10.48550/arxiv.2408.04667.
229. Ma, W.; Yang, C.; Kästner, C. (Why) Is My Prompt Getting Worse? Rethinking Regression Testing for Evolving LLM APIs. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI* **2024**, pp. 166–171. https://doi.org/10.1145/3644815.3644950.
230. Dulac-Arnold, G.; Levine, N.; Mankowitz, D.; Li, J.; Paduraru, C.; Gowal, S.; Hester, T. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning* **2021**, *110*, 2419–2468. https://doi.org/10.1007/s10994-021-05961-4.
231. Reddy, P.; Gujral, A. EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System. *arXiv (Cornell University)* **2025**, *abs/2509.10540*. https://doi.org/10.48550/arxiv.2509.10540.
232. Elad, B. AI in Finance Statistics By Market Size, Adoption, Use Cases, Trends and Facts (2025). https://electroiq.com/stats/ai-in-finance-statistics/, 2025.
233. AI, M. Achieving AI ROI: Key Findings from the 2025 Forbes AI Study. https://www.mavvrik.ai/forbes-ai-study-2025/, 2025.
234. Anthropic. Claude Code by Anthropic | AI Coding Agent, Terminal, IDE, 2025.
235. OpenAI. Codex | AI Coding Partner from OpenAI, 2025.
236. Anysphere. Cursor: The best way to code with AI, 2023.
237. Infinite, A. WisPaper - Your AI Scholar Search Engine | AI Literature Review, 2025.
238. Jimenez, C.E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; Narasimhan, K.R. SWE-bench: Can Language Models Resolve Real-world Github Issues? In Proceedings of the The Twelfth International Conference on Learning Representations, 2024.
239. Zhou, S.; Xu, F.F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. WebArena: A Realistic Web Environment for Building Autonomous Agents. In Proceedings of the The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024.
240. Hughes, C. The Agentic AI Governance Blind Spot. https://www.resilientcyber.io/p/the-agentic-ai-governance-blind-spot, 2026.
241. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. *CoRR* **2023**, *abs/2302.12173*, [2302.12173]. https://doi.org/10.48550/ARXIV.2302.12173.
242. Lazer, S.J.; Aryal, K.; Gupta, M.; Bertino, E. A Survey of Agentic AI and Cybersecurity: Challenges, Opportunities and Use-case Prototypes. *CoRR* **2026**, *abs/2601.05293*, [2601.05293]. https://doi.org/10.48550/ARXIV.2601.05293.
243. Booth, H.; Fisher, W.; Galluzzo, R.; Roberts, J. Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization. Concept paper, National Institute of Standards and Technology, 2026.
244. Infocomm Media Development Authority. Model AI Governance Framework for Agentic AI. Technical report, Infocomm Media Development Authority, Singapore, 2026.
245. Staufer, L.; Feng, K.; Wei, K.; Bailey, L.; Duan, Y.; Yang, M.; Ozisik, A.P.; Casper, S.; Kolt, N. The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems. *CoRR* **2026**, *abs/2602.17753*, [2602.17753]. https://doi.org/10.48550/ARXIV.2602.17753.
246. Bresnahan, T.F.; Greenstein, S. Technological Competition and the Structure of the Computer Industry. *The Journal of Industrial Economics* **1999**, *47*, 1–40, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-6451.00088]. https://doi.org/https://doi.org/10.1111/1467-6451.00088.
247. Jovanovic, B.; Rousseau, P.L. General Purpose Technologies. Working Paper 11093, National Bureau of Economic Research, 2005. https://doi.org/10.3386/w11093.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2026 doi:10.20944/preprints202608.1248.v1

248. Helpman, E.; Trajtenberg, M. A Time to Sow and a Time to Reap: Growth Based on General Purpose Technologies. NBER Working Paper 4854, National Bureau of Economic Research, 1994. Available at SSRN: https://ssrn.com/abstract=226530.
249. Acemoglu, D.; Autor, D. Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings.; Elsevier, 2011; Vol. 4, *Handbook of Labor Economics*, pp. 1043–1171. https://doi.org/https://doi.org/10.1016/S0169-7218(11)02410-5.
250. Autor, D.H.; Levy, F.; Murnane, R.J. The Skill Content of Recent Technological Change: An Empirical Exploration*. *The Quarterly Journal of Economics* **2003**, *118*, 1279–1333, [https://academic.oup.com/qje/article-pdf/118/4/1279/5427313/118-4-1279.pdf]. https://doi.org/10.1162/003355303322552801.
251. Acemoglu, D.; Restrepo, P. Tasks, Automation, and the Rise in U.S. Wage Inequality. *Econometrica* **2022**, *90*, 1973–2016, [https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19815]. https://doi.org/https://doi.org/10.3982/ECTA19815.
252. Aghion, P.; Howitt, P. A Model of Growth Through Creative Destruction. *Econometrica* **1992**, *60*, 323–351.
253. Babina, T.; Fedyk, A.; He, A.; Hodson, J. Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics* **2024**, *151*, 103745. https://doi.org/https://doi.org/10.1016/j.jfineco.2023.103745.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.