Useful Agentic AI: A Systems Outlook

Melissa Pan[β], Yuxuan Zhu[𝕌], Jared Quincy Davis[σ*], Riccardo Cogo[‡]
Lakshya A Agrawal, Negar Arabzadeh, Xiaoyuan Liu, Huanzhi Mao, Sid Pallerla, Tianneng Shi, Alexander Xiong [β], Alessandro Basile[‡], Emmanuele Lacavalla[‡], Shuyi Yang[‡], Diana Arroyo, Paul Castro[†], Daniel Kang[𝕌], Joseph Gonzalez, Koushik Sen, Dawn Song, Ion Stoica, Matei Zaharia[β], Marquita Ellis[†]
[β]UC Berkeley, [σ]Stanford University, [𝕌]UIUC, [*]Mithril, [‡]Intesa Sanpaolo, [†]IBM Research
ai-agent-survey@googlegroups.com

Abstract

To effectively meet and anticipate the requirements of practically impactful (useful) Agentic AI systems, including but not limited to new metrics, benchmarks, programming and runtime systems, it is necessary to first understand the use-cases and state of progress (or lack thereof). We present our perspective from practitioner interviews and questionnaire responses, critically supplementing the wider body of published data (papers, blogs, OSS). We aggregated and distilled this data for 10+ cases spanning enterprise, software, and scientific domains, with several more in progress. Our aim is to bridge this academic-industry divide, steering our own and other's systems research toward the most impactful problems.

Though Agentic AI innovations are rapidly proliferating across academia and industry, the requirements and key challenges between academic and enterprise settings differ significantly. Academic literature is currently skewed toward pushing capabilities of LLM-based agents and Artificial General Intelligence. Industry use-cases rely on data and models of multiple modalities, focus on specific enterprise-grade versus general functionality, and cannot ignore latencies, cost, scalability, security, and many other practical deployment concerns. We find, while not yet useful for all tasks, Agentic AI systems are becoming increasingly useful for facilitating, augmenting, and/or accelerating scoped manual and software tasks. Our principle findings is that Useful Agentic AI systems currently rely on a combination of human and machine verifiers, with the most sophisticated (autonomous) systems supported by strong verifiers from mathematics, computer and computational sciences. The lack of strong verifiers appears to contribute more to the lag between industry and academia than a lack of AI capabilities. Revisiting complexity analysis and systems for verification is a critical and timely contribution area for interdisciplinary and industry-academia collaboration.

Keywords

AI Agents, Agentic AI, Agentic AI Systems, Survey

1 Introduction

Since generative AI began unlocking advanced capabilities for AI Agents and general Compound AI Systems [Zaharia et al. 2024], building Agentic AI systems has become the focus of hundreds of startups [CBInsights 2025], thousands of academic publications, and truly massive open online courses [Song and Chen 2024; Song et al. 2025]. Academic and industry researchers have prototyped new capabilities for everything from multi-agent tutoring [Schmucker et al. 2024], to physical asset management [Timms et al. 2024], to self-driving science and engineering [Juraj Gottweis 2025; Novikov et al. 2025]. However, whether and to what extent AI Agents are practically useful is an open debate [Challapally et al. 2025]. Many capabilities demonstrated in academic spheres have yet to make an appearance in production systems. Problems solved and unsolved in industry for securing, scaling, and deploying production Agentic AI systems remain largely unpublished.

A few recent publications with similar motivation, offer lengthy analyses without broader concrete use-case analysis [Krishnan 2025; Liu et al. 2025] or limit scope to evaluation systems [Yehudai et al. 2025] or usability [Shome et al. 2025]. We take the position that reliable, deployable Agentic AI systems, like their wider family of Compound AI Systems [Zaharia et al. 2024], require not only demonstration of AI capability but also robust supporting systems integration. Further, that the most impactful system innovation will follow from analyzing the application use-cases first.

Hence, this study targets industry case studies, seeking to understand potential impact and progress (or lack thereof) across use-cases and domains, focusing on the relatively recent onset of foundation-model-infused AI agents. In light of this focus and industry cases incorporating multi-modal, hybrid, and foundation models broadly, we do not restrict the study to purely LLM-based agents for which there are numerous surveys of academic progress [Guo et al. 2024; Wang et al. 2024]. For case study selection, we prioritize production-grade cases self-described as "AI agent(s)" or "Agentic AI system(s)", rather than restricting selection according to one of the many (academic) definitions of "AI agent". We re-derive from cross-comparison what is considered successful (or unsuccessful) –useful– Agentic AI in practice.

Our central questions are, what is actually working (and not) in practice from a computational systems perspective? Why, what is making Agentic AI *useful* or not and to what extent? Is driving down inference latency the only or most important contribution systems researchers can make? Where can academic and industry research communities –and systems and AI research communities– better align to realize the potential of Agentic AI in practice?

In contrast to academic literature, we find deployed (useful) Agentic AI systems commonly focus on relatively well-defined, simple tasks repeatedly executed by human customers or employees. These include information retrieval, static and semi-static workflow execution across application domains. Successful cases reduce time-to-completion (increase operational throughput), and/or lower knowledge and skill requirements for completing tasks involving multiple system interfaces and domain-specific knowledge and procedures. Reduction in time-to-completion is commonly measured relative to human time to complete the same task, yielding a wide spectrum of use-case-derived latency requirements. Hence, the range of contribution opportunities for the systems community includes exploiting relaxed latency constraints in scheduling and resource management, and is not limited to single-inference latency minimization. Further, we find useful Agentic AI systems rely on a combination of human and machine verifiers, with the most sophisticated (autonomous) systems supported by strong verifiers from mathematical, computer and computational sciences. The lack of strong verifiers appears to contribute more to the lag between industry and academia than a lack of AI capabilities. Revisiting complexity analysis and systems for verification is a critical and timely contribution area for interdisciplinary and industry-academia collaboration.

2 Methodology

We found openly available case study data, such as published literature, blog posts, open-source code repositories, present an incomplete picture of the landscape for systems research and development. To ground and expand openly available data, we engaged industry participants through interviews and questionnaires.

Interviewers were selected to maintain organizational neutrality, assigned roles, and completed a series of pre-, post-, and in-interview procedures. The structure of interviews was determined by a preset list of 11 topic groups (below), and the availability of respective answers from open sources that need only be verified via interview.

(1) **The root problem (benefit) the system is addressing (providing):** What is the ultimate benefit? What is the system replacing and why?

(2) **Key success metrics and evaluation mechanism:** What tools, techniques, systems, etc. are used to ensure the system meets user and stakeholder objectives? Is data corresponding to the expected or past system behavior available for the evaluation?

(3) **Key aspects of the system design and implementation:** What programming framework was used? What is the general architecture? What are the steps, stages, and cycles? How are common components (e.g. routers, LLM-as-a-Judge, other verifiers, HIL) combined and why? What is the ratio of automation to human interaction and why—by design or limitation?

(4) **The state of the system or its development:** Is the system in production, or was it never meant for production (purely for AI research, learning, upskilling)? Was the system prototyped for production but abandoned—why, and what were the critical limitations?

(5) **Known constraints or requirements of end-users and stakeholders:** What are the security, regulatory, SLO/SLA requirements?

(6) **Advantage of an agentic AI system solution over alternative approaches:** what is the advantage in your view?

(7) **System dependencies and complexity:** what is the quantity, quality, and availability of tools, verifiers, data, etc.?

(8) **End-user quantity, expertise levels, and organizational domains.**

(9) **Estimated cost versus value or benefit.** Including sunk and ongoing costs of developing and operating the system versus its estimated value.

(10) **System stakeholders:** Who ultimately benefits from deployment? Who is impacted by safety, security, etc. failures and limitations?

(11) **Your role and activities:** What is your involvement with the agentic AI system(s) you are describing?

For breadth, understanding how far-reaching cross-case-study observations were, we iteratively crafted a questionnaire for mass distribution. The questionnaire was distributed to technical groups across the AI Alliance Agents-in-Production Meetup [1], the Berkeley RDI Agentic AI Summit [2], and collaborators' professional networks. Our questionnaire design informed by interviews, seeks to avoid response priming, facilitate downstream quantitative analysis, and facilitate broad participation by restricting the length, terminology, and the technical-depth and disclosure-depth necessary to complete the questionnaire. Further, we agreed upfront to aggregate and anonymize all data. As an aside, interviews revealed that what is considered confidentially-innovative in the space of Agentic AI varied significantly across organizations. In summary, our study integrates openly available data with perspectives from industry practitioners to answer, what is *useful* Agentic AI?

3 Use-Case Data

Table 1: Anonymized in-depth case study descriptions.

Business Operations
Insurance claims workflow automation
Customer care internal operations assistance
Human resources information retrieval and task assistance
Communications (U.S. and Latin America)
Automotive communication services
Communication automation services
Scientific Discovery
Biomedical sciences workflow automation
Software & Business Operations
Data analysis for enterprise
Enterprise cloud engineer and business assistance
Site reliability incident diagnoses and resolution
Software products question answering
Software DevOps
Spark version code and runtime migration
Software development life cycle assistance end-to-end

Data for the study combines public data with 12 in-depth case studies (Table 1). Summary findings are presented in aggregate per confidentiality agreements with the sources. These 12 were selected based on originators' availability for interviews, application-diversity, and development status preferring those in production (total 8) or pre-production piloting (4). The cases spanned business

[1] ttps://luma.com/x16vikh7. Note: last accessed: 6 Oct. 2025.

[2] https://rdi.berkeley.edu/events/agentic-ai-summit. Note: last accessed: 6 Oct. 2025.

Table 2: Case Counts by Source Company Characteristics

Stage		Continents		Countries		Case Use	
Mature	6	1	5	One	5	External	8
Late	1	2 to 4	4	Tens	6	Internal	4
Growth	1	5 to 6	3	Hundreds	1		
Early	3						
Seed	1						

operations (3), software development and operations (2), a tightly integrated combination thereof (4), scientific discovery (1), and enterprise communication services (2). They also differed in their intended use, 4 targeting internal software and business operations, and 8 targeting external (enterprise) consumers. Table 2 additionally lists statistics on the spread of case sources by company stage, continent and country spread. The questionnaire response data – over 400 responses and growing – is used complementarily for broader validation.

4 Summary Findings

Agentic AI is being applied to human-facing and highly interdisciplinary problems. The tasks and correctness conditions lack formal specification [Stoica et al. 2024]. The performance or evaluation metrics and systems do not always match the ultimate goal or benefit of the system. Still, some things are working (in production). Others are not. The following unpacks these observations from industry case studies complemented with mass survey results and public data.

4.1 HIL and Fundamental Complexity

Across software development, enterprise and scientific discovery systems, we are observing Agentic AI being applied to tasks specified by humans. This was true of every single case in our study and 94.8% direct consumers of survey respondents' systems. The common goal of these systems is to reduce the human time necessary for completing scientific, software, legal, or business processes. Human interfaces designed to delight were a secondary requirement for productivity rather than the goal or primary requirement. An important exception however included cases governed by policies requiring human oversight of Agentic AI systems, such as the EU Artificial Intelligence Act. For these cases, the engineers designed with HIL in mind from the start. In general, AI coding agents are among if not the most advanced Agentic AI systems in production, but still exhibit a spectrum of HIL rather than absence of HIL. At one extreme, state-of-the-art (SoA) systems claim multiple hours of autonomous execution of delegated development tasks (e.g. up to 30 hrs [Anthropic 2025]). However, human input is commonly still required for major actions such as PR approval [GitHub 2025]. On the other extreme, human approval of fine-grained code modifications (e.g. in-IDE auto-completion) is still ubiquitous. SoA IDEs even provide options for varying the level of Agentic AI autonomy [Cursor Team 2025; Deshmukh et al. 2025; GitHub 2025].

We have not encountered a production-track Agentic AI system implemented without human-in-the-loop (HIL), even with the scope narrowed to business, software engineering, and scientific applications. Figure 1 introduces terminology and illustrates our observations of how production-track systems split functionality between HIL and automated methods across the key runtime stages of task specification, solution generation, solution verification, and ongoing evaluation. The categories we focus on here to illuminate technical opportunities are HIL in verification and evaluation.

Applied Agentic AI raises the level of abstraction for the fundamental compute unit, but is not yet asking nor answering the questions complexity theory enabled computer scientists to answer to date. Classical complexity theory enabled asking fundamental questions, *is a solution computable in the first place? In how much time (our lifetime)? Is there a difference between the generation versus verification time and space complexity?* It enabled answering these questions prior to building impractical solutions, and it led to constructive solutions (approximations) even for *hard* problems. We do not have the equivalent scaffolding for Agentic AI systems.

4.2 HIL in Verification and Evaluation

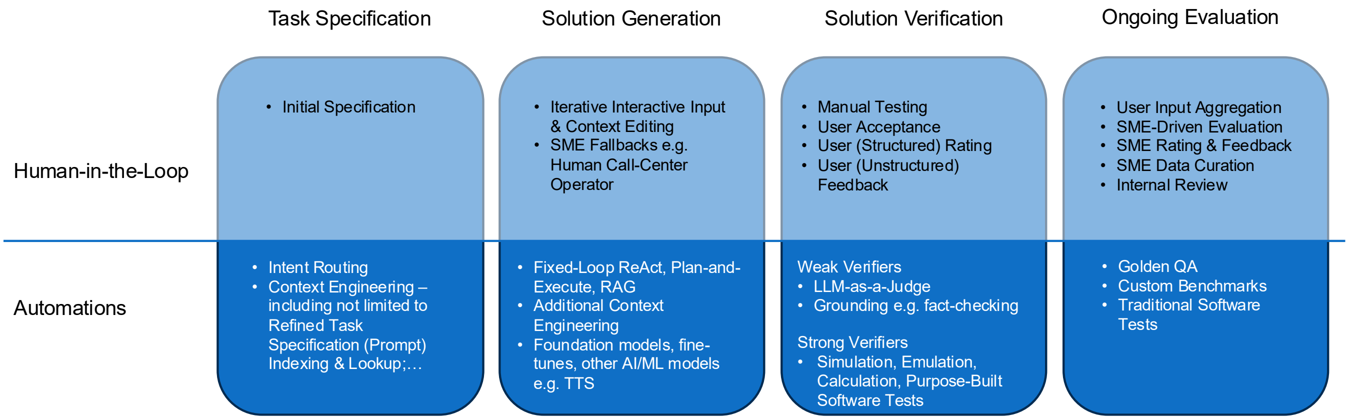
Verification mechanisms measure whether the systems output meets the correctness conditions for a specific task, problem, or input, and may be triggered online or during offline evaluation. Evaluation mechanisms measure how well the system performs on given collections of tasks or inputs and metrics over time.

Beyond software development, practitioners are recognizing and capitalizing on the tractability of the decision (action) space for known human workflows across domains, from human support systems to scientific lab routines. However, automation for recognizing (verifying) correct workflow completion is lacking in many cases. The Subject Matter Expert (SME) knowledge has not been captured to a sufficient degree in data or tooling. Hence, while the data is being captured and tooling created, deployed systems are forced to rely on end-user and/or SME verification.

The SME may no longer be the solution generator (completing the workflow manually) but they are still in the pipeline, at the verification and evaluation stages (Figure 1). We found the pervasiveness and extent of human SME involvement across case studies surprising, given attention across academia and industry to evaluation. Capitalization [Foody 2025] is indicative of a broader trend. Opportunities are still open to develop (reusable) domain-specific evaluation methods, and to improve and scale evaluation data ingestion, curation, synthetic generation, and so on.

We have also observed Agentic AI system architectures in deployment are relatively simple compared to SoA academic systems. Simple fixed-loop ReAct-based [Yao et al. 2023] and RAG-based [Lewis et al. 2020] agents are common to an extreme. Clear benefits are software maintainability leading to higher reliability. Non-technical factors (e.g. education, business decisions, etc.) are of course at play. As for technical factors, given the sophistication gap between production-grade Agentic AI systems for mathematical and scientific applications and other applications with other evidence, we suspect lack of strong verifiers is key.

Verifiability implies repeatability. In practice, we have observed test-and-retry to be a common motif in production systems. Advanced production systems commonly employ sandboxing, emulation, or simulation of Agentic AI outputs during runtime and/or optimization.



	Task Specification	Solution Generation	Solution Verification	Ongoing Evaluation
Human-in-the-Loop	• Initial Specification	• Iterative Interactive Input & Context Editing • SME Fallbacks e.g. Human Call-Center Operator	• Manual Testing • User Acceptance • User (Structured) Rating • User (Unstructured) Feedback	• User Input Aggregation • SME-Driven Evaluation • SME Rating & Feedback • SME Data Curation • Internal Review
Automations	• Intent Routing • Context Engineering – including not limited to Refined Task Specification (Prompt) Indexing & Lookup;…	• Fixed-Loop ReAct, Plan-and-Execute, RAG • Additional Context Engineering • Foundation models, fine-tunes, other AI/ML models e.g. TTS	Weak Verifiers • LLM-as-a-Judge • Grounding e.g. fact-checking Strong Verifiers • Simulation, Emulation, Calculation, Purpose-Built Software Tests	• Golden QA • Custom Benchmarks • Traditional Software Tests

Figure 1: Use-cases in practice split the stages of task specification, generation, verification, and ongoing evaluation between human-in-the-loop and automated methods. Lists in each stage are example methods observed in production-track use-cases, and do not comprehensively represent the plethora of methods in the literature.

4.3 HIL ≠ Human-Attention Span Latency

We observed that human-in-the-loop did not directly translate human-attention-span latency requirements, a multi-decade standard [Nielsen 1993]. The case studies revealed instead a spectrum of latency requirements. At the low-latency end of the spectrum, we found real-time applications such as speech-driven customer service targeting order 100*ms* latencies. At the other end of the spectrum, we encountered a "longer is better" motif, in which "thinking" or execution time indicated runtime robustness and higher expected output quality (higher confidences). Public code-related products exemplify this well [Anthropic 2025; Deshmukh et al. 2025].

In between these extremes, latency targets are evolving with user preferences and task complexity. Respectively, production development efforts are starting to rely more on new usage data collection, curation, and analysis (notably, with LLMs in the loop) than on best-practices pre-dating the current generation of Agentic AI systems. Futhermore, practitioners' comparisons of the agentic system's latencies to the previous (manual) system latencies show delightful speedups. It might take the system 15 minutes but human response turnaround is several hours — not due to the complexity of the task but due the limited availability of human hours for e.g. 24/7 operations. Operations across sectors are finding Agentic AI systems a useful fast-lane for customer and employee response.

Overall, instead of hardware-based latency targets or even past standards, latency trade-offs are being evaluated with respect to human time and availability. If the system takes days, weeks, or months to complete an essential task but it takes a human potentially decades [Novikov et al. 2025] to complete that same task, the system is useful. Bringing the time for single-version software migration down to months from approx. 2 years is likewise *useful*. In the same vein, if it takes a human Subject Matter Expert seconds to complete a task (e.g. code line completion, customer service response) the latency targets we have observed are closer but not necessarily less than that human benchmark due to the availability of human hours. Understanding the use-cases helps us understand the optimization opportunities.

4.4 Limitations

This study is inherently limited by access to production system data. Our sampling is relatively small but already fruitful for industry-academia gap analyses. Data collection is ongoing and will no doubt lead to report revisions. We hope the study fuels discussions on broader collaboration and information exchange between industry and academia.

5 Conclusions and Outlook

Large Language Models (LLMs) sparked an irresistible opportunity to again attempt replacing software interfaces with natural language across applications – web applications, data analytics, search, HR and enterprise systems, and many others. Existing user interfaces have insufficiently reduced complexity, complexity stemming from quantity, quality, or (dis)organization of information or the complexity of multi-system operations with multi-modal I/O. However, LLMs alone are insufficient for industry tasks — tasks requiring multiple modalities of structured and unstructured data, data not limited to text, and procedures and correctness criteria encoded in human cognitive versus digital formats.

Agentic AI systems composed of foundation models and prior generations of models and tools [Zaharia et al. 2024] are starting to bridge the gap and exceed language interface application scopes, interfacing between human-machine and machine-machine environments to simplify and automate operations across business, science, technology, and beyond. The innovations are so impressive and fast-paced, it is difficult to distinguish short-lived demos from stable progress and gauge which contribution areas will have impact. We studied production-grade, deployment-track Agentic AI and dubbed them *useful* Agentic AI systems.

Based on our findings, a useful Agentic AI system is not an Agentic AI system capable of the most complex and general tasks. A useful Agentic AI system improves the ratio of simplistic, computable tasks allocated to machines to (relatively) complex, interesting, rewarding tasks allocated directly or indirectly to humans. Increasing the task complexity machines are able to process requires addressing not only generation complexity but also verification complexity.

There is a significant lag between agentic AI techniques in industry versus academia. While academia continues to push AI capabilities generally, subdomains showing the least lag in industry are founded on fields with strong verifiers –mathematical, computational, and scientific fields. This seems to imply lack of verification is (in part) contributing more to broader industry lag than lack of AI capabilities. Even for such leading applications, there is recognition that the key to solving more complex problems with Agentic AI is bolstering verification [Cornelio et al. 2025; Midha 2025].

Related but not limited to verification, areas for researchers (across disciplines) include scalable evaluation data ingestion, curation, synthetic generation; sandboxing, emulation, simulation environments for AI agents; machine-to-machine agentic AI; formalizing use-case characterization and goodput metrics; fast general intent and task routing; and many others. We plan to present more insights following the final rounds of data collection and analysis.

Acknowledgments

We thank our many anonymous participants without whom beginning this study would not be possible.

References

Anthropic. 2025. Claude Sonnet 4.5 maintained autonomous focus for over 30 hours. https://www.anthropic.com/news/claude-sonnet-4-5. Last Accessed: 2025-10-02.

CBInsights. 2025. *The AI agent market map.* Technical Report. CB Information Services, Inc. https://www.cbinsights.com/research/ai-agent-market-map/ Last accessed 10 July 2025..

Aditya Challapally, Chris Pease, Ramesh Raskar, and Pradyumna Chari. 2025. *The GenAI Divide: State of AI in Business 2025.* Technical Report. MLQ.ai and Project NANDA. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf Preliminary findings from AI implementation research.

Cristina Cornelio, Takuya Ito, Ryan Cory-Wright, Sanjeeb Dash, and Lior Horesh. 2025. The Need for Verification in AI-Driven Scientific Discovery. arXiv:2509.01398 [cs.AI] https://arxiv.org/abs/2509.01398

Cursor Team. 2025. Cursor IDE: AI-Powered Development with Rule-Based Control. https://cursor.com/docs. Last Accessed October 4, 2025.

Neeraj Deshmukh, Siddhant Pardeshi, Brian Elliott, Jack Blundin, Advika Sadineni, Yash Bolishetti, David Rome, Simon Mead, and Advait Sadineni. 2025. Blitzy System 2 AI Platform: Topping SWE-bench Verified. https://paper.blitzy.com/blitzy_system_2_ai_platform_topping_swe_bench_verified.pdf. Last Accessed: 2025-10-04.

Brendan Foody. 2025. Welcome to the Era of Evals. https://mercor.com/blog/welcome-to-the-era-of-evals/. Last Accessed: 2025-10-03.

GitHub. 2025. GitHub Copilot Agent executes multi-step tasks autonomously via GitHub Actions. https://github.com/newsroom/press-releases/coding-agent-for-github-copilot. Last Accessed: 2025-10-02.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

Alexander Daryin Tao Tu Anil Palepu Petar Sirkovic Artiom Myaskovsky Felix Weissenberger Keran Rong Ryutaro Tanno Khaled Saab Dan Popovici Jacob Blum Fan Zhang Katherine Chou Avinatan Hassidim Burak Gokturk Amin Vahdat Pushmeet Kohli Yossi Matias Andrew Carroll Kavita Kulkarni Nenad Tomasev Vikram Dhillon Eeshit Dhaval Vaishnav Byron Lee Tiago R D Costa José R Penadés Gary Peltz Yunhan Xu Annalisa Pawlosky Alan Karthikesalingam Vivek Natarajan Juraj Gottweis, Wei-Hung Weng. 2025. Towards an AI co-scientist. https://storage.googleapis.com/coscientist_paper/ai_coscientist.pdf.

Naveen Krishnan. 2025. AI Agents: Evolution, Architecture, and Real-World Applications. arXiv:2503.12687 [cs.AI] https://arxiv.org/abs/2503.12687

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, and Sebastian Riedel. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS).* https://arxiv.org/abs/2005.11401

Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, Yuheng Cheng, Suyuchen Wang, Xiaoqiang Wang, Yuyu Luo, Haibo Jin, Peiyan Zhang, Ollie Liu, Jiaqi Chen, Huan Zhang, Zhaoyang Yu, Haochen Shi, Boyan Li, Dekun Wu, Fengwei Teng, Xiaojun Jia, Jiawei Xu, Jinyu Xiang, Yizhang Lin, Tianming Liu, Tongliang Liu, Yu Su, Huan Sun, Glen Berseth, Jianyun Nie, Ian Foster, Logan Ward, Qingyun Wu, Yu Gu, Mingchen Zhuge, Xiangru Tang, Haohan Wang, Jiaxuan You, Chi Wang, Jian Pei, Qiang Yang, Xiaoliang Qi, and Chenglin Wu. 2025. Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems. arXiv:2504.01990 [cs.AI] https://arxiv.org/abs/2504.01990

Anjney Midha. 2025. *Investing in Periodic Labs.* https://a16z.com/announcement/investing-in-periodic-labs/ Last Accessed: 2025-10-04.

Jakob Nielsen. 1993. *Usability Engineering.* Academic Press, Inc., Harcourt Brace Company, San Diego, USA.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131* (2025).

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. RuffleRiley: Insights from Designing and Evaluating a Large Language Model-Based Conversational Tutoring System. arXiv:2404.17460 [cs.CL] https://arxiv.org/abs/2404.17460

Pradyumna Shome, Sashreek Krishnan, and Sauvik Das. 2025. Why Johnny Can't Use Agents: Industry Aspirations vs. User Realities with AI Agent Software. arXiv:2509.14528 [cs.HC] https://arxiv.org/abs/2509.14528

Dawn Song and Xinyun Chen. 2024. Large Language Model Agents. https://llmagents-learning.org/f24. Last access: 2025-07-10.

Dawn Song, Xinyun Chen, and Kaiyu Yang. 2025. Advanced Large Language Model Agents. https://llmagents-learning.org/sp25. Last access: 2025-07-10.

Ion Stoica, Matei Zaharia, Joseph Gonzalez, Ken Goldberg, Hao Zhang, Anastasios Angelopoulos, Shishir G Patil, Lingjiao Chen, Wei-Lin Chiang, and Jared Q Davis. 2024. Specifications: The missing link to making the development of LLM systems an engineering discipline. *arXiv preprint arXiv:2412.05299* (2024).

Alexander Timms, Abigail Langbridge, and Fearghal O'Donncha. 2024. Agentic Anomaly Detection for Shipping. In *NeurIPS 2024 Workshop on Open-World Agents.*

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/2210.03629

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on Evaluation of LLM-based Agents. arXiv:2503.16416 [cs.AI] https://arxiv.org/abs/2503.16416

Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, and Ali Ghodsi. 2024. The Shift from Models to Compound AI Systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

accepted 3 October 2025