AI Agents for Actuarial: Or How I Learned to Stop Worrying and Love the Bot

An Introduction to AI Agents for Actuaries

DANIEL CRAIG RAMSAY

APR 22, 2025

Share

1. Introduction

In 2025, the use of generative artificial intelligence (GenAI) is no longer a distant possibility in the actuarial profession—it has firmly arrived, with research tools like ChatGPT and AI coding assistants like Github Copilot becoming widely used in day-to-day tasks. However, a lesser known but potentially more transformative area in the field of AI that is gaining traction is the concept of AI agents. In this article, which will be the first in a series regarding the topic of AI agents, we explain what agents are, how they work, their potential use cases and why their emergence is so significant for the profession.

The actuarial field is broad and multidisciplinary, traditionally involving data analysis, model development, risk assessment, and strategic decision-making, among many other tasks that demand advanced financial acumen. While the field has seen automation before, it has largely relied on specialized software tools used in tandem with manual processes. AI agents, however, go a step further, offering the potential to perform high-level tasks that previously were beyond the capabilities of rule-based processes.

As these autonomous systems gain ground, they bring a mix of opportunities and concerns. On one hand, automating routine tasks allows actuaries to focus on strategic decision-making and risk assessment. On the other, the opacity and potential unreliability of AI-driven models raise questions about regulatory compliance, model transparency, and governance. These factors are especially important in highly regulated sectors such as insurance and financial services, where accuracy and accountability are paramount. In this series, we will also explore how actuaries can harness the power of AI agents while safeguarding professional standards and the financial interests of consumers..

2. What Is an AI Agent?

Historically, the term “agent” in artificial intelligence denoted any system that could autonomously perceive its environment via sensors and enact changes via effectors (Russell & Norvig, 2016). Early work in distributed AI and expert systems relied on deterministic or probabilistic rule-based approaches (Lesser & Corkill, 1978), offering only limited adaptive behavior. (Wright, R., 2020) noted that these early agents were largely theoretical or considered niche applications.

However, a major inflection point occurred with the development of the transformer architecture (Vaswani et al., 2017), triggering rapid improvements in both natural language processing and the generation of robust, coherent outputs from large-scale models. This ushered in a new era wherein agents, powered by Large Language Models (LLMs), could adjust their actions, and gradually evolve beyond static pre-programmed responses.

There are many modern definitions of GenAI-based agents in circulation. The following two, from two leading AI labs, may help clarify the concept.

OpenAI (2024) definition:
"Agents represent systems that intelligently accomplish tasks, ranging from executing simple workflows to pursuing complex, open-ended objectives."

Anthropic (2024) Definition:
Anthropic defines AI agents along a spectrum:

- Workflows: Systems where large language models (LLMs) and tools are orchestrated through predefined code paths.
- Agents: Systems where LLMs dynamically direct their own processes and utilize tools autonomously.

Other definitions also include:

- **IBM** (2025) defines an AI agent as “a system or program that autonomously performs tasks by designing its workflow, leveraging large language models to interpret user instructions, and executing required subtasks without continuous human oversight.”
- **Oracle** (2025) describes them as “software entities that examine their environments, set goals, plan actions, and adapt based on new data.”

- **The World Economic Forum** (2024) characterises AI agents as systems that detect and respond to environmental cues to achieve outcomes, while also necessitating robust governance and risk-management frameworks.

It is important to note that the term "AI agent" has become somewhat nebulous. In commercial circles, the term is used to describe anything from LLM-based chatbots to fully autonomous systems equipped with sophisticated decision-making capabilities. For clarity, this series focuses on **software-based agents that operate on computers that are capable of planning multi-step tasks and tool-use.** This is opposed to the already well-understood concept of AI chatbots, which we define as those systems which primarily only generate text-based responses and do not have the capability to dynamically interact with the real world.

3. Achieving Agency: Planning, Memory and Tool Use

3.1. Limitations of Co-pilot Systems and why planning and tool use is important

Since the release of ChatGPT in November 2022, we have continued to see incremental improvements in the intelligence and capabilities of AI. Each passing month seems to usher in more intelligent, more reliable and more capable systems.

In a professional setting, the immediate benefits of using LLMs have been observed in areas such as coding, document summarization, and data extraction—tasks historically time-consuming for knowledge workers. For actuaries, these productivity gains translate into faster data validation, automated model documentation, and the ability to replicate complex spreadsheet calculations efficiently. In these contexts, LLMs serve as "co-pilots" that augment human expertise rather than replace it.

Despite impressive performance in text generation, most standard LLM-based chatbots lack the capability to execute multi-step tasks or autonomously invoke external functions, typically supporting only turn-based responses. The transition to agentic systems requires additional structures—such as planning processes and tool integration —that enable these models to engage in iterative and dynamic problem solving.

A more concrete example of the limitations of purely LLM-based systems would be assigning the task to an AI of building an actuarial model from scratch. Creating an actuarial model does not just involve writing code, but also:

- loading and cleansing data;
- inputting, compiling and executing code within modelling platforms;
- testing and validating results;
- generating analysis reports;
- and documentation of models.

Each of these tasks will require the AI to have the ability to interact with software, i.e., **tool use.** Moreover, in order for the AI to approach the task holistically, rather than tackling each task piecemeal, the AI will need to be able to orchestrate the steps and adapt to unforeseen circumstances that inevitably arise during model development.

This orchestration requires **planning** capabilities that go beyond the capabilities of early LLMs and only now is becoming within reach of the most capable frontier models. For an AI to plan effectively, it must be able to develop a coherent strategy, remember and monitor progress, identify when adjustments are needed, and seamlessly transition between different tools and processes. Without these planning capabilities, even the most advanced LLM would struggle to maintain the coherence and continuity necessary for complex, multi-stage tasks like actuarial modeling.

3.2 Tool Use: Extending Capabilities beyond Text Generation

Tool use in AI agents typically involves invoking APIs, executing code, or even interacting with graphical user interfaces. A good example is OpenAI's Assistant Code Interpreter (OpenAI, 2025) which is able to load files and execute code directly within a dedicated virtual environment. Similarly, tools like Auto-GPT, OpenAI Operator, and Claude Computer Use demonstrate how LLMs can be integrated with tool-calling frameworks to autonomously navigate file systems, retrieve web data, or even run Python scripts.

Effective tool use requires two main properties: the ability to parse and interpret tool specifications and generate tool calls in structured formats (such as JSON or XML). Knowing what is the right tool to use and when is a skill can be second nature to humans, but for an AI this can be a difficult challenge that can cause even the most sophisticated models to stumble. Indeed, AI models capable of tool-use often need to be fine-tuned specifically for this task to ensure they meet a suitable level of reliability and accuracy.

3.3 Planning to plan

Planning is a cornerstone of autonomous behaviour in AI agents and there are many ways to achieve competency in this area, namely through techniques such as prompt engineering and flow engineering as well as through the advanced reflective capabilities of reasoning models and multi-agent systems. Moreover, to support planning, it is crucial for the AI to have a reliable system of memory which is able to scale with the complexity of the task and allow the AI agent to monitor its progress in achieving its goals.

3.3.1 Memory and Retrieval in AI Agents

A critical component enabling effective planning and tool use in AI agents is memory. Modern agents typically employ multiple memory layers to maintain context, learn from past actions, and adapt to new situations much like how humans process and retain information.

The design of memory architecture is a broad and quickly evolving field. However, CrewAI represents a notable example of an advanced memory architecture implementation. Its system consists of:

- Short-Term Memory using Retrieval-Augmented Generation (RAG) to quickly recall recent interactions.
- Long-Term Memory that preserves valuable task outcomes and leverages a SQLite3 database for structured storage.
- Entity Memory, which organizes data about encountered entities using RAG as its backbone.
- Contextual Memory, which fuses different memory layers to ensure continuity across complex tasks.
- User Memory, which supports personalization by recording specific user inputs and preferences (CrewAI Docs, 2024).

By default, CrewAI uses OpenAI embeddings to represent memory, with data files managed via robust file storage directories adjustable through environment variables. More details can be found in their documentation (https://docs.crewai.com/concepts/memory).

3.3.2 Chain-of-Thought (CoT)

CoT prompting is one such prompt engineering technique, as introduced by Wei et al. (2022) which improves the reasoning capabilities of LLMs by guiding the model through intermediate logical deductions before arriving at a final answer. This has been shown to dramatically improve the accuracy and reliability of LLM output due to the AI needing to make shorter inferential leaps, reducing the risk of making mistakes.

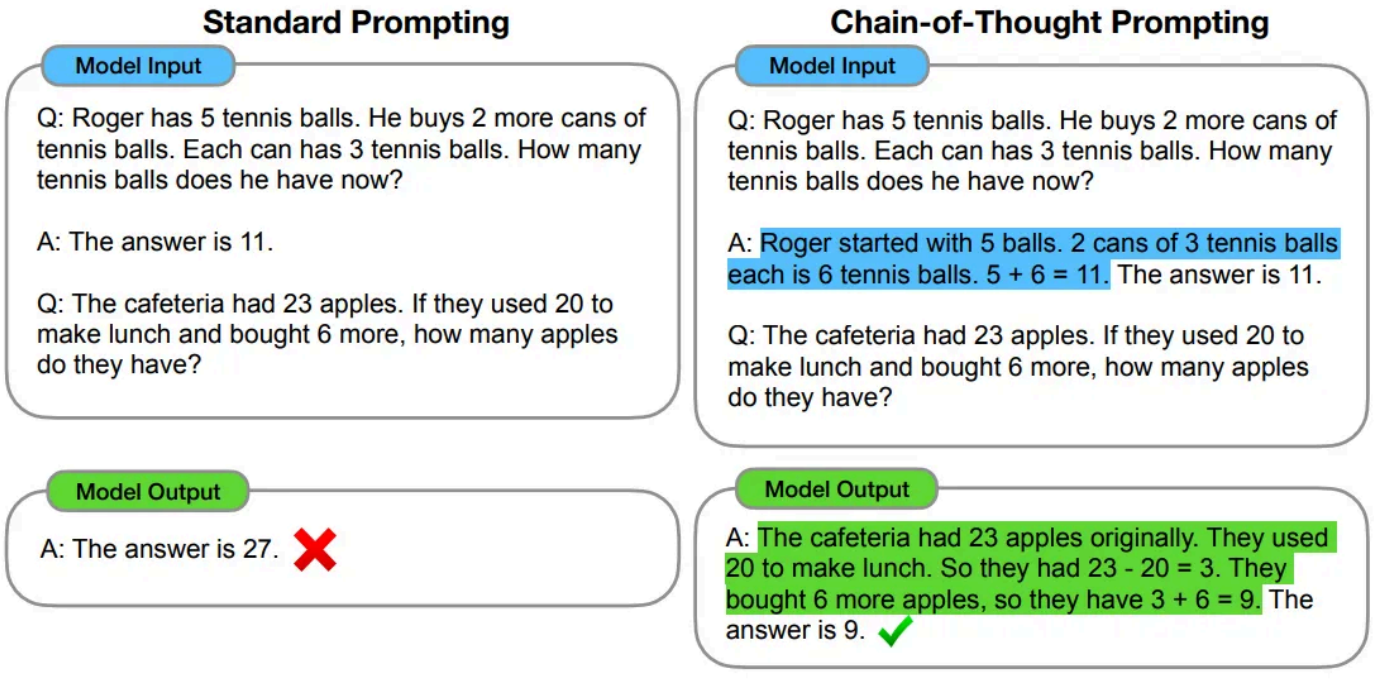

Figure 3.3.2: Generating working before providing a final answer can significantly improve the accuracy of LLM output. Source: Wei et al. (2022).

3.3.3 Flow engineering

This is an alternative more explicit approach to planning that further structures the LLM planning process by pre-defining sequences of actions or planning heuristics in system prompts and conversation management while maintaining the agent's ability to dynamically adjust its approach. Flow engineering techniques can orchestrate multiple iterations of hypothesis generation, tool invocation, and error correction that are crucial for tasks requiring long-term planning and adaptability.

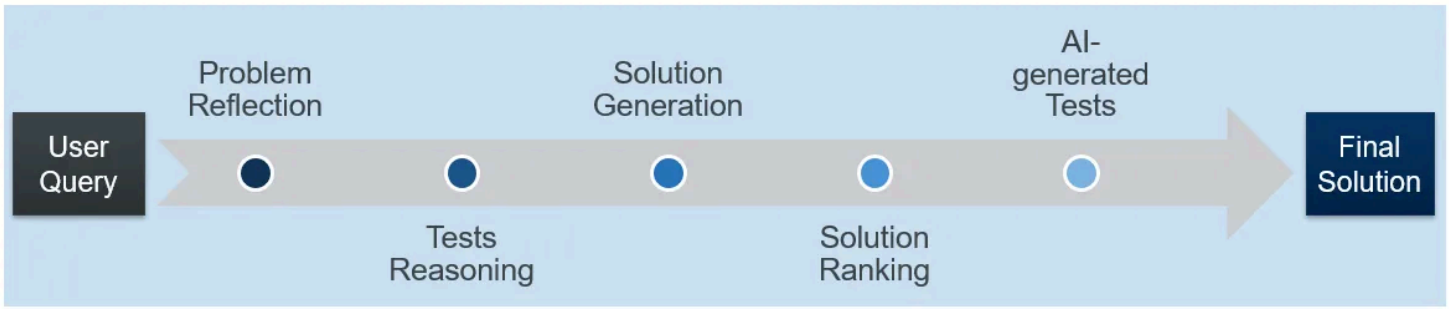

Figure 3.3.3: Rather than generating a solution in a single-turn, the process of identifying a solution is broke up into a structured series of subtasks across a multi-turn conversation.

3.3.4 Reasoning models

Reasoning models, like OpenAI's o1, leverage chain-of-thought approaches being embedded directly into the behavior of the model, which enables them to outperform traditional LLMs in complex tasks requiring significant planning or working. Unlike standard models that generate immediate responses, reasoning models produce internal reasoning tokens—iterative steps that break problems into smaller subtasks, reducing the risk of errors arising from large inferential jumps.

Moreover, reasoning models can evaluate multiple possible solutions, weigh evidence, and identify potential flaws in their own reasoning through self-reflection. This self-reflective capability enables iterative refinement before producing a final answer. By internally simulating different reasoning paths and evaluating their validity, these models can tackle tasks requiring strategic planning, causal reasoning, and complex decision-making with significantly improved performance.

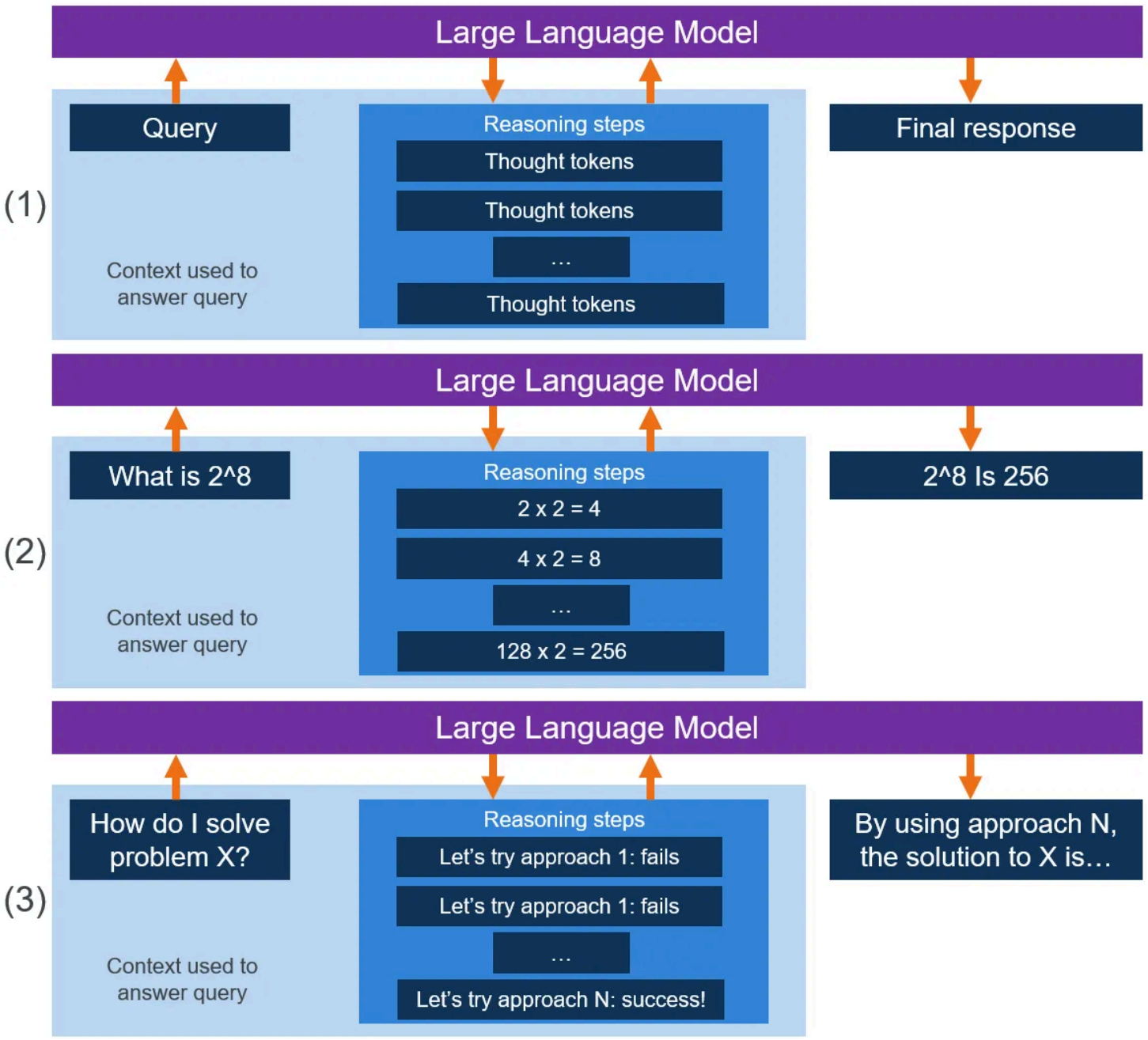

Figure 3.3.4: (1) Rather than relying on memorisation, reasoning models allow conclusions to be derived using more granular steps. Reasoning tokens can be used to break down problems into more manageable sub-problems (2) or can be used to self-reflect on different approaches to solving a problem (3).

3.3.5 Multi-Agent Systems: Delegation and Hierarchical Collaboration as the Next Abstraction

Beyond single-agent systems, multi-agent systems (MAS) represent a new paradigm in creating robust, scalable, and flexible AI. MAS leverages the strengths of multiple specialised agents that can collaborate or independently handle subtasks of a larger objective.

Multi-agent architectures typically involve defining multiple highly specialized roles with clearly delineated responsibilities. These systems may also include a central manager agent which orchestrates the process by identifying and evaluating tasks then assigning them to the most appropriate agent. These systems are particularly promising for applications in dynamic, real-world settings where a single large monolithic agent may struggle to meet all the requirements of a complex process that has a large number of rules to follow.

For example, in an actuarial modelling project, one agent might manage data retrieval, while another focuses on risk analysis, and yet another generates documentation. This division of labour not only enhances accuracy but also enables parallel processing that can drive down task completion times.

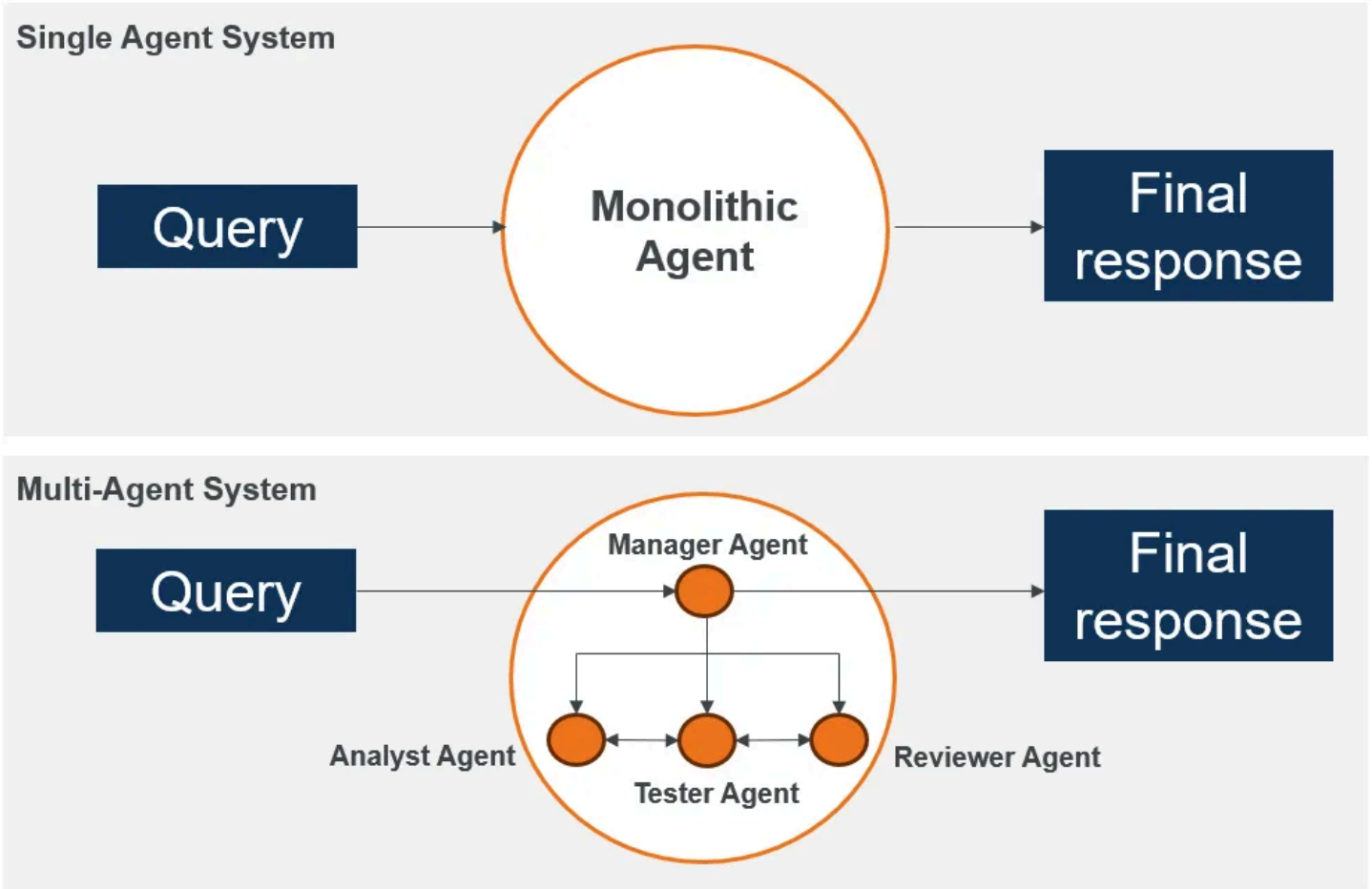

Figure 3.3.5: A single agent system will have one agent that is responsible for all tasks while a multi-agent system will be able to delegate subtasks to specialised agents.

4. Current capabilities and the rate of improvement

As AI agents mature, it is essential to evaluate their performance in real-world tasks. Several noteworthy benchmarks have emerged:

- **OSWorld** is a unified, scalable benchmark for evaluating multimodal agents on authentic computer tasks in real operating system environments. Derived from genuine human computer interactions, it spans a wide range of applications—from file management and office productivity to multi-app workflows—across systems like Ubuntu and Windows. OSWorld's execution-based evaluation leverages raw keyboard and mouse actions together with screenshots and accessibility tree data, enabling reproducible assessments of an agent's ability to perform complex, open-ended tasks. Notably, while human users achieve around 72% task success, leading AI agents scored only about 12% at time of publication, highlighting significant challenges in GUI grounding and operational reasoning (OSWorld Benchmark, 2024).
- **WebVoyager** is a benchmark for evaluating web agents powered by Large Multimodal Models (LMMs). It assesses the ability of agents to perform end-to-end tasks by interacting with real-world websites. The benchmark is derived from 15 popular websites and includes tasks such as searching for products and information, answering questions based on academic papers, performing calculations and simplifications among others. WebVoyager utilizes screenshots and textual content of interactive elements as input for the agents. On the WebVoyager benchmark, the WebVoyager agent achieves a 59.1% task success rate, compared to GPT-4 (All Tools) scores 30.8% and 40.1% for a text-only version of WebVoyage (WebVoyager, 2024).

Recent evaluations have measured the performance of advanced computer-using agents:

- Claude Computer Use scored 56% on WebVoyager and around 22% on OSWorld.
- OpenAI Operator (CUA) has demonstrated an 87% success rate on WebVoyager and 38.1% on OSWorld.

Another key benchmark to follow is SWE-bench which is designed to evaluate AI models' abilities to autonomously resolve software engineering issues which have been sourced from real-world GitHub projects. It provides a structured framework where AI agents are tasked with interpreting problem statements and proposing code changes to address specific issues.

When SWE-bench was launched in March 2024, AI models equipped with Retrieval-Augmented Generation (RAG) capabilities achieved only single-digit success rates. Notably, GPT-3.5, the model that ignited the public's fascination with GenAI upon the release of ChatGPT in November 2022, managed a success rate of just 0.40%. However, the landscape has rapidly evolved. Early AI frameworks like SWE-agent saw success rates in the low double digits. With the advent of reasoning models such as OpenAI O1, O3, and DeepSeek R1, performance has soared. OpenAI O3, released in December 2024, achieved an astonishing 71.7% success rate on SWE-bench Verified, marking a significant leap from previous benchmarks. This rapid progress within less than a year underscores the tremendous advancements in AI technology .Although not yet widely integrated into software development workflows, AI agents' exceptional debugging capabilities will soon make them an indispensable part of the standard software development toolkit.

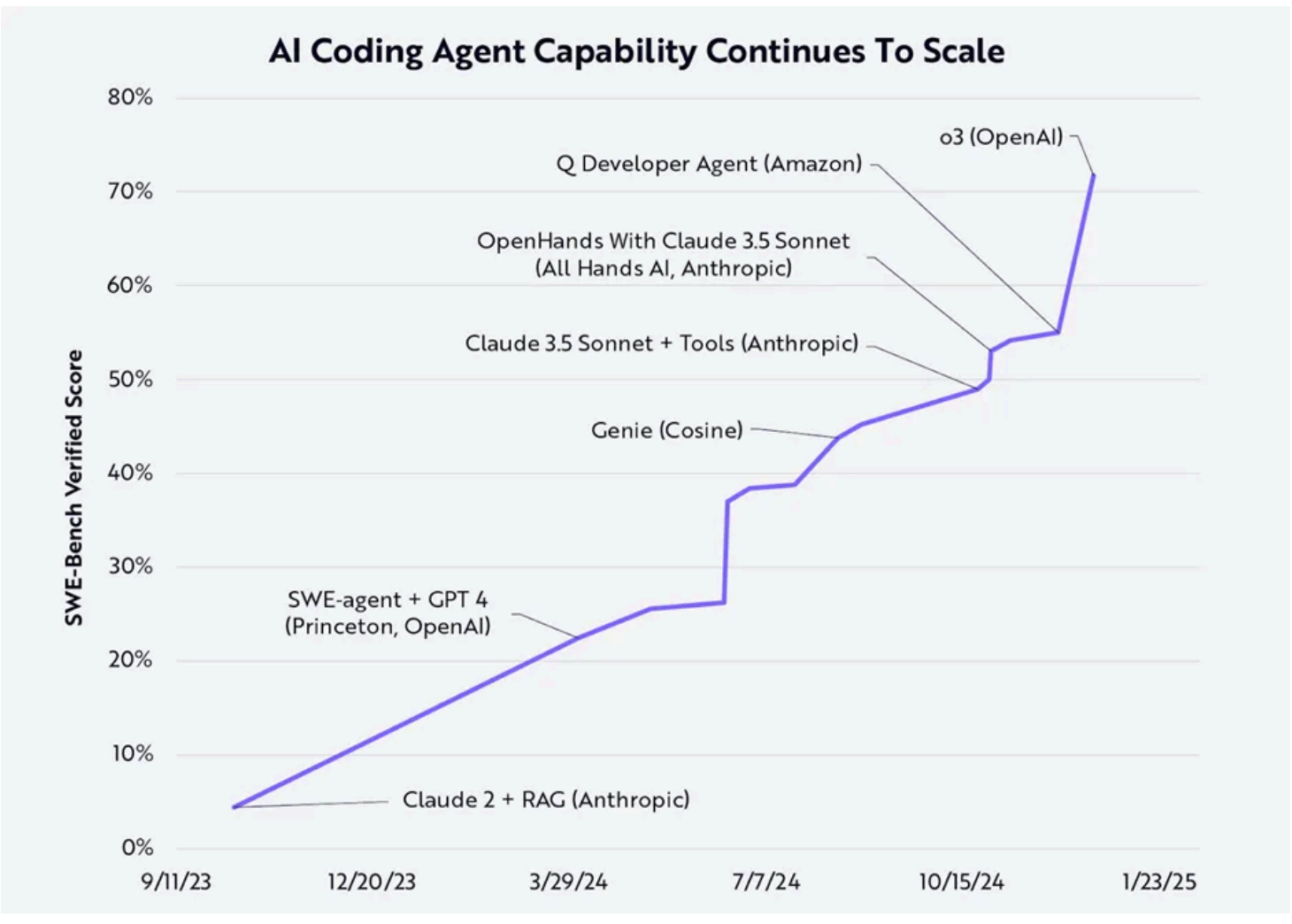

Figure 4: Progress on the SWE-Bench Verified over time. Source: ARK Invest (2024).

One might ask the valid question of what is the relevance these benchmarks have to actuarial work. First, one must consider that actuaries are knowledge workers, which

means we spend a significant amount of time using a computer performing analytical work. The continuing improvement in the OSWorld and WebVoyager benches illustrate that AI is gaining a broad level of competency in use of computers that underlie the analytical work which comprise a proportion of daily activities of actuaries.

To fully comprehend the implications of SWE-bench, one must recognize the parallels between the modeling work of actuaries and software development. Both fields demand a high level of programming knowledge and involve managing complex systems to perform their respective roles effectively. It is therefore, not much of a leap of the imagination to envisage that in the near future, we will begin to see the same gains in the capabilities of AI agents to perform actuarial modelling work as has been seen in the software developments (This topic will be covered in more depth in the second article of this series).

The point we aim to illustrate to the reader is that the trend lines are clear: these models are becoming increasingly capable over time, performing ever more complex tasks, and this progress shows no sign of slowing down. Hence as we go into the next decade we could see the role of actuaries transform dramatically as agents encroach on more and more of the activities which make up the typical work of an actuary.

5. How AI Agents are Likely to Impact Actuarial Work

The introduction of autonomous AI agents into actuarial work promises to overhaul traditional workflows. By integrating AI agents capable of autonomous planning, external tool use, and dynamic adaptation, many routine tasks could be automated. In this section, we examine two areas in detail where AI is likely to lead to significant productivity gains, namely administrative tasks and modeling.

5.1. Administrative Task Automation

Actuaries, like many knowledge workers, spend a significant portion of their time on administrative tasks that, while necessary, divert attention from higher-value analytical work. AI agents offer promising solutions to streamline these activities. The administrative applications below represent early opportunities for AI agent adoption that deliver clear time-savings while posing minimal risk to core actuarial functions. Some of these capabilities are already available within existing software, making them accessible entry points for actuarial teams looking to begin their AI agent deployment journey.

5.1.1 Document Management and Reporting

Current Process: A typical actuarial team might spend a large number of hours per month preparing standard reports, involving data extraction, formatting spreadsheets, creating visualizations, and drafting explanatory text.

AI Agent Solution: An autonomous reporting agent could:

- Monitor data repositories for updates
- Extract relevant figures and trends
- Generate standardized reports with appropriate commentary

5.1.2 Meeting Coordination and Follow-up

Current Process: Actuarial teams often coordinate multiple stakeholder meetings requiring preparation of materials, note-taking, and action item tracking.

AI Agent Solution: Meeting assistant agents could:

- Schedule meetings
- Prepare agendas by analysing previous discussions and current priorities
- Record and transcribe meeting content
- Extract action items and automatically assign them in project management tools
- Generate meeting summaries with highlighted decisions and follow-up requirements

5.1.3 Regulatory Compliance Documentation

Current Process: Maintaining documentation for regulatory compliance often involves manually updating numerous documents when assumptions or methodologies change.

AI Agent Solution: A compliance documentation agent could:

- Monitor changes to models, assumptions, or data sources
- Automatically update all affected documentation
- Flag potential regulatory concerns
- Maintain a comprehensive audit trail of all changes

5.1.4 Email and Communication Management

Current Process: Actuaries spend a significant fraction of their workday managing emails and communications, triaging requests, and searching for information.

AI Agent Solution: Email management agents could:

- Categorise and prioritize incoming communications
- Draft responses to routine inquiries
- Extract action items from message threads
- Compile relevant background information for complex requests
- Manage follow-ups and reminders for outstanding items

5.2. Case Study: Modeling Process Transformation

While administrative tasks offer immediate opportunities, the transformation of core modeling processes represents a more profound shift in how actuarial work is

performed. Rather than focusing on the technical aspects which will be covered in the second article of this series, this case study examines how AI agents might reshape the modeling workflow and team dynamics.

5.2.1 Continuous Assumption Updating

Current Process: Most actuarial teams conduct formal assumption reviews on fixed schedules (quarterly or annually), requiring significant manual effort to gather market data, analyse experience, and build supporting analyses.

AI Agent Solution: An assumptions monitoring agent could:

- Continuously monitor relevant data sources and market indicators
- Generate preliminary analysis of the impact on existing models
- Prepare draft assumption updates with supporting evidence
- Alert actuaries when human review is warranted

5.2.2 Model Version Control and Governance

Current Process: Maintaining version control across complex actuarial models often relies on manual documentation processes and individual discipline, leading to inconsistencies and governance challenges.

AI Agent Solution: A model governance agent could:

- Automatically document all model changes with contextual information
- Compare versions to identify material changes in methodology or outputs
- Generate impact assessments when models are modified
- Enforce governance workflows requiring appropriate approvals
- Maintain comprehensive model inventories with metadata

5.2.3 Cross-functional Model Integration

Current Process: Integrating models across actuarial, finance, and risk functions often involves manual processes, data transfers, and reconciliation efforts.

AI Agent Solution: Integration agents could:

- Maintain mappings between different models
- Automatically transform data between required formats
- Identify and flag inconsistencies between departmental assumptions

5.2.4 Facilitating Collaboration

Current Process: Complex modeling solutions often require collaboration between highly specialised individuals, with knowledge sharing constrained by availability and awareness of expertise.

AI Agent Solution: An agent that facilitates collaboration could:

- Connect modeling teams with relevant internal experts based on problem characteristics
- Extract and share relevant modeling approaches from previous projects
- Generate adaptive documentation that explains model components at varying technical levels

5.3. Implementation Considerations

For actuarial teams considering AI agent adoption in the future, several key factors would need to be considered when determining a suitable implementation strategy:

- **Start with bounded tasks:** Initial deployments should focus on well-defined, lower-risk tasks with clear evaluation criteria. AI agents will start performing simple repetitive tasks long before they will be deployed to more abstract or strategic work.
- **Hybrid workflows:** For the foreseeable future, it is clear we still require a human-in-the-loop for the production of any material piece of work, process design should consider how best to facilitate this.
- **Address data privacy early:** Many actuarial processes involve sensitive data requiring proper safeguards and controls. The same precautions and processes that came about with the advent of GDPR legislation will still apply to workflows which have AI agent systems interacting with sensitive data.
- **Focus on explainability:** Any output generated by an AI agent should be accompanied by commentary and justification explaining why the output was created. This ensures that these systems do not function as a black box.

It is likely in the future, progress in deploying agentic systems will start with administrative use cases, which will build familiarity and trust before we see these models encroaching on more specialised modeling applications. As of April 2025, the ChatGPT moment for AI agents in the enterprise is yet to occur, however that time is likely not far away.

6. A Future with AI Agents:

It is the authors' view that AI agents are on the cusp of becoming a new general-purpose technology like steam power, electricity, or the internet. Actuaries, known for their aptitude in forward-thinking and risk assessment, will appreciate the need to focus on broader trend lines rather than today's headlines. If benchmark improvements in AI capabilities continue to be as reliable as technology trends such as Moore's law, we may soon find ourselves in a world where the marginal cost of intelligence approaches zero. In this new era, an actuary's role must be redefined to ensure enduring relevance—focusing on high-level oversight, expert interpretation of analytical results, and strategic insights that require a uniquely human touch.

By embracing AI agents thoughtfully, actuaries can position themselves at the intersection of human expertise and artificial intelligence. The future of our field belongs not to those who resist this evolution but to those who adapt, innovate, and

leverage these new capabilities to solve increasingly complex problems in risk management, insurance and pensions spheres. As we navigate this transition, the actuarial profession has a unique opportunity to redefine itself and emerge stronger, more efficient, and more valuable than ever before.

7. What's Next in Our AI Agent Research Series

This article forms the first part of a series exploring the multifaceted impact of AI agents on actuarial modelling and broader enterprise applications. Future instalments will delve deeper into modelling, the impact of AI on actuarial careers, governance and other topics.

References

- Anthropic. (2024). *Building effective agents.* Retrieved from https://www.anthropic.com/research/building-effective-agents
- Auto-GPT. (2023). *GitHub repo.* Retrieved from https://github.com/Significant-Gravitas/AutoGPT
- IBM. (2025). *AI agents.* Retrieved from https://www.ibm.com/think/topics/ai-agents
- Lesser, V. R., & Corkill, D. D. (1978). The Distributed Vehicle Monitoring Testbed: A tool for investigating distributed problem-solving networks. *AI Magazine*, *4*(3), 15.
- OpenAI. (2020). *GPT-3: Language models are few-shot learners.* arXiv preprint arXiv:2005.14165.
- OpenAI. (2024). *Agents.* Retrieved from https://platform.openai.com/docs/guides/agents
- OpenAI. (2025). *Code interpreter.* Retrieved from https://platform.openai.com/docs/assistants/tools/code-interpreter
- Oracle. (2025). *AI agents.* Retrieved from https://www.oracle.com/uk/artificial-intelligence/ai-agents/
- OSWorld Benchmark. (2024). *OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments.* arXiv preprint arXiv:2404.07972.
- Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- SWE-bench. (2024). *SWE-bench: Can language models resolve real-world GitHub issues?* arXiv preprint arXiv:2310.06770.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000-6010).
- WebVoyager Benchmark. (2024). *WebVoyager: Building an end-to-end web agent with large multimodal models.* arXiv preprint arXiv:2401.13919.

- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*. arXiv preprint arXiv:2201.11903.
- Russell, S. J., & Norvig, P. (2020). Artificial intelligence: A modern approach (4th ed.). Pearson.
- World Economic Forum. (2024). *Navigating the AI frontier: A primer on the evolution and impact of AI agents.* Retrieved from https://reports.weforum.org/docs/WEF_Navigating_the_AI_Frontier_2024.pdf.

Discussion about this post

Comments Restacks

Write a comment...

© 2025 Daniel Ramsay · Privacy · Terms · Collection notice
Substack is the home for great culture