AI Explained · AI Agents and Tool Use, Part 1

What Are You Asking the Agent to Do?

By Dr Peter Fields

THE 2-MINUTE SUMMARY

An AI agent can use tools and decide what to do next in light of what it finds. How much discretion it has depends on the task and the software around it.

Agents can help with work such as checking contracts or answering routine customer queries. Time spent checking the work, and what the supplier actually charges for, can make a difference to whether it pays.

If you may change providers, keep your instructions, source documents and tests somewhere you control. Test the new agent on the awkward cases as well as the easy ones.

A PDF with a gap in it

When I asked an AI tool with access to my files to prepare the downloadable PDFs for *Images, Voice and Video*, I spotted a large gap before the copyright section. I asked it to check the other PDF for the same problem and keep the source links clickable. I described what I wanted to see on the page. Working out what had caused the gap was part of the job.

A chatbot could tell me where to look in Word. A tool could export the document to PDF. A fixed workflow could export it, count the pages and check the links in a set order. An agent could inspect the result, decide what to check next and, if the layout needed changing, make a further adjustment and export the document again. The difference is how much of the method I leave to the system.[1]

Tools let the system do things such as read a file, search a database or create a document. The model proposes an action. The application decides whether the action is available, runs it if allowed and returns the result. The model can then use that result to decide what to try next.[1]

Why are we seeing more agents now?

The idea has been around for a while. In 2023, Toolformer showed that a language model could learn when to ask a calculator or a search tool for help. ReAct showed how a model could use what came back to decide on another action. Some of those experiments used simulated tasks. They help explain the method behind agents you can use today.[2]

Two things made it easier to turn the idea into a product. Models got better at completing longer sequences of actions in tests, and software makers gave them more ways to work with files and applications. Anthropic's public computer-use beta in October 2024 could interact with a screen, though even scrolling and dragging could be unreliable. In January 2025, OpenAI released a preview of Operator, which could work in a web browser.[3,4]

Geek Box · What does a two-hour benchmark mean?

In a 2025 METR study, OpenAI's o3 had a 50 per cent time horizon of about 110 minutes. That means the test predicts success about half the time on tasks that would take a skilled person about 110 minutes. The researchers found that this measure had roughly doubled every seven months for leading models in their historical data. If you want 80 per cent reliability, the benchmark points to shorter tasks. The tests were mainly software tasks. The clock measures the human's time.[4]

Who is using agents, and for what?

In law, Harvey describes how SUSE's legal team uses AI to review agreements against standard positions. Its account of an acquisition describes lawyers checking the flagged issues against the underlying documents. In investment research, Bridgewater's Aaron Linsky said Claude powered early versions of an assistant that could work with code and charts. These accounts describe particular uses reported by the suppliers and their customers.[5]

Salesforce describes how Wiley piloted an agent for routine customer queries, including password resets, freeing staff to deal with more complicated cases. GitHub's 2026 Octoverse report said that, by September, more than 40 per cent of proposed code changes submitted for review on GitHub started with a detectable first draft from an agent. How much checking those proposals needed is a separate question.[6]

At home, you might ask an agent to sort copies of your household bills by date and list any months for which it can find no bill. Some consumer tools can keep files and instructions together across tasks, so you can return to the same job when a new bill comes in.[7]

What are you paying for?

A subscription may include a usage allowance. A developer may pay for the amount of material the model reads and generates, with separate charges for some tools. A business may be charged for actions, conversations or results. Under Intercom's published rules, a customer who leaves after receiving an answer can count as an assumed resolution. If you pay by result, read the definition. Suppliers could increasingly combine these charges, with a monthly fee for access and separate charges for heavier jobs.[8]

There is also your time. In an early 2025 METR study, a small group of experienced open-source developers took 19 per cent longer when AI tools were available. Results from a later experiment pointed towards time savings. In February 2026, METR explained why the size of the saving was hard to pin down. Some developers avoided submitting work they were unwilling to do without AI, and others did different work while their agents ran. For your own trial, count the time spent setting up the agent, checking the result and putting mistakes right.[9]

Could I take my agent with me?

Suppose a firm uses an agent to review supplier contracts. It checks each agreement against the firm's checklist, points to the clauses behind its findings, flags exceptions and prepares a note for a solicitor.

The solicitor decides what to send to the client. What would the firm need to change if it moved to a different provider?

Keep the checklist, review instructions, templates and source documents in systems the firm controls, along with any records it needs of completed work and approvals. If the firm built the workflow itself, it may be able to change the model while keeping most of that workflow in place. A move between hosted products can mean rebuilding document search, reconnecting accounts and setting up the approval steps again.[10]

Two open specifications can help. The Model Context Protocol (MCP), introduced by Anthropic and contributed to a Linux Foundation project in 2025, lets compatible products connect to the same tools. Open Responses, announced by OpenAI in 2026, gives compatible software a common way to call a model. Check what you can take out of the old product, too. The consumer exports for ChatGPT and Claude include chat history. Claude's individual export excludes the original uploaded files. A firm should keep its own copies.[10,11]

Before switching, test both systems on sample agreements the firm is entitled to share. Choose cases where you know what should happen, including a missing schedule and a draft that must wait for approval. Before the new system takes over any scheduled work, turn off the old schedule and account for jobs already finished or still pending.

What is a sensible first task?

The UK's National Cyber Security Centre recommends starting with a tightly bounded trial. Choose a small job with a clear end and a result you can check. With household bills, give the agent copies from a period you've already checked. See whether it spots the same gaps and what it tells you when the evidence is incomplete.[12]

My PDF example gave the agent room to decide what to try next. The consequences would have been different if it had also been able to replace the download on my website. Part 2 starts with what happened when an agent working on an app could reach the live data.

Sources and notes

1. Anthropic (2024), Building Effective AI Agents, sections on workflows and agents. Alan Chan, Rebecca Salganik, Alva Markelius and colleagues (2023), Harms from Increasingly Agentic Algorithmic Systems, §2.1. Anthropic · Chan et al.

2. Timo Schick, Jane Dwivedi-Yu, Roberto Dessì and colleagues (2023), Toolformer: Language Models Can Teach Themselves to Use Tools, §2. Shunyu Yao, Jeffrey Zhao, Dian Yu and colleagues (2023), ReAct: Synergizing Reasoning and Acting in Language Models, §§2–4. Toolformer · ReAct

3. Anthropic (2024), Introducing Computer Use. OpenAI (23 January 2025), Introducing Operator. Computer use · Operator

4. Thomas Kwa, Ben West, Joel Becker and colleagues, Measuring AI Ability to Complete Long Software Tasks, v4, §§2–4. See also METR's explanation of the time-horizon measure. Paper · METR charts and methods

5. Harvey, SUSE customer story, sections on contract review and due diligence. Anthropic, Claude for Financial Services, Bridgewater's account of its Investment Analyst Assistant. SUSE · Bridgewater

6. Salesforce, Wiley customer story, pilot section. GitHub, 2026 Octoverse report, September figures and glossary on agentic pull requests. Wiley · Octoverse

7. Anthropic, Organize your tasks with projects in Claude Cowork. The household example is illustrative. Support article

8. OpenAI, About ChatGPT Pro Tiers and API Pricing. Salesforce, Agentforce Pricing. Intercom, Fin AI Agent Outcomes, section on assumed resolution. Charging examples checked 19 November 2026. Subscription · API · Salesforce · Intercom

9. Joel Becker, Nate Rush, Beth Barnes and David Rein (2025), Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, abstract and §§2–3. METR (February 2026), follow-up. Study · Follow-up

10. Model Context Protocol, Architecture Overview. Linux Foundation (9 December 2025), announcement of the Agentic AI Foundation. OpenAI (January 2026), API changelog. Open Responses, overview and specification. MCP architecture · Linux Foundation · OpenAI changelog · Open Responses

11. OpenAI, Exporting your ChatGPT history and data. Anthropic, Export your Claude data, opening description. ChatGPT export · Claude export

12. UK National Cyber Security Centre (15 May 2026), Thinking carefully before adopting agentic AI. NCSC guidance