What happened when we asked AI agents to help someone move house

9th November 2026 - Tom Loosemore plus lots of help from Claude Opus5.5

Why bother testing?

After asking various AI agents what they might find useful, I've been toying with the idea of a change of address directory for personal AI agents. It would be one trustworthy place that tells your agent which public services, and which companies, need to know you've moved, and how to tell each of them.

I'd always assumed the hard part would be gathering the data and keeping it fresh. Then a nagging doubt set in. The real risk might be persuading agents to bother with it at all, when their default is to search the web. If search already gets them there, why would they use anything else?

So before building anything, I thought it wise to find out how good their default way of find this stuff out actually is. That meant testing it.

How the tests worked

Think of it as setting an exam, with the answers marked by someone who isn't the student.

1. **Invent some people who are moving.** Each one gets a short request, written as a real person might type it to their agent. For example: "I'm moving from my rented flat in Manchester to Edinburgh on 1 December. I get PIP, I work part time, and I don't have a car. Who do I need to tell, and how?"
2. **Write the model answer.** For each move, list what a complete answer must include. We researched every line from the organisation's own website (gov.uk, council sites, the bank's or insurer's help pages) and noted how sure we were. We also added a few traps: common mistakes a good answer shouldn't make, such as "you need to update your passport". You don't.
3. **Ask the agent.** Each request goes to the agent with nothing extra: no hints, and no mention of the directory idea.
4. **Mark the answer.** A separate AI model compares the answer with the key, line by line. It notes what was found, what was half-found, which traps were fallen into, and anything sensible the agent suggested that our key had missed.

The headline score is simple: **what share of the things that really matter did the agent find?** By "really matter" I mean a legal duty, or something where you'd lose money or support if you forgot.

Before trusting the marker, we checked it, and scored the answers. We also timed each answer and, where we could, estimated what it cost for the agent’s time and effort.

The scenarios we tested

We used the same eight ‘house moves’ twice. The first time, the question was about government and council services. The second time it was about companies and charities. Each move is set on 1 December 2026.

Who	From	To	What makes it tricky
Renter with a car and a parking permit	Hackney	Tower Hamlets	Two London boroughs, two parking schemes
Renter with a car	Colchester	Cardiff	England to Wales
Renter on PIP (a disability benefit)	Manchester	Edinburgh	England to Scotland, where PIP is replaced
Homeowner who drives	Bristol	Belfast	England to Northern Ireland: different rates, licences and suppliers
Renter who drives	Leeds	Leeds	Moving within the same council
Pensioner (helped by her son)	Norwich	Truro, Cornwall	A county and a district council at one end, a single council at the other
Family on Universal Credit, two children at school	Birmingham	Coventry	Leaving a council tenancy, new school places
Company director who drives	Cambridge	Oxford	Companies House, and parking run by the county

That adds up to 15 places, 16 councils, and every UK nation.

For **government**, the model answers listed 77 things to do. They ranged from contacting DVLA, HMRC and DWP to Scotland's new disability benefit, Northern Ireland's rates office and a garden waste bin in Colchester.

For **companies and charities**, each person named their providers, chosen from 20 organisations:

- **Banks:** Barclays, HSBC, Lloyds, Nationwide, Monzo
- **Insurers:** Admiral, Aviva, Direct Line
- **Pensions:** Nest, Legal & General
- **Energy:** Octopus, British Gas
- **Broadband:** BT, Virgin Media, Sky
- **Mobile:** EE, Vodafone, O2
- **Charities:** the National Trust and the RNLI

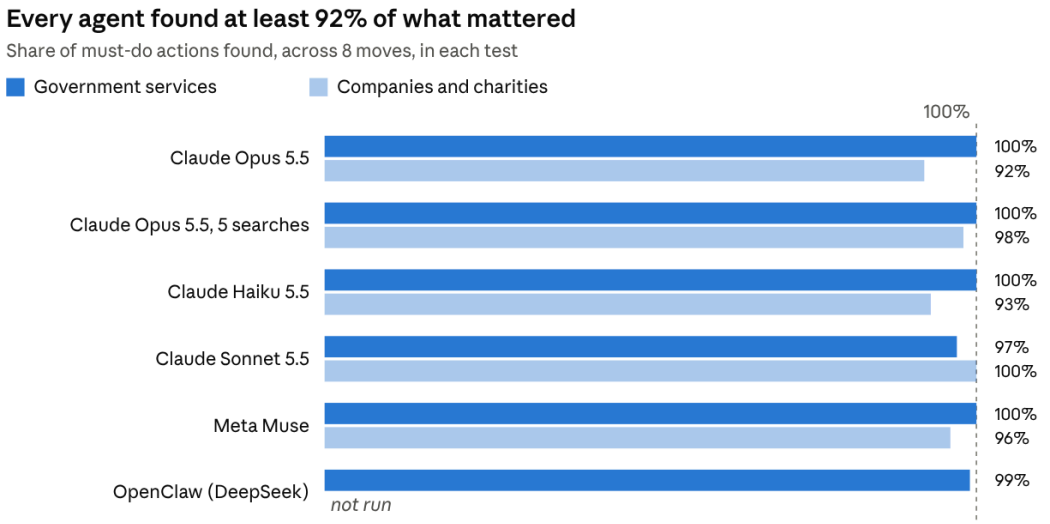
There, the test was less about who to tell (they'd been told) and more about how: the right route, the notice needed, and the catches.

Who took the test, and how they did

Six agents took part:

- **Four versions of Anthropic's Claude,** from the most capable (Opus) to the cheapest and quickest (Haiku). One Opus was allowed only five web searches, to see whether thrift hurts.
- **Meta's Muse,** a personal agent, which I drove through its website.
- **OpenClaw,** an open-source agent I run myself, here using a model called DeepSeek.

I'd have liked to include Instinct and Claude Duet too, but they were too much of a faff to automate.



Our tests, 8–9 November 2026 · 8 house moves per test

Government services: better than I expected

Every agent found at least 97% of the things that really mattered, even the agent that barely searched. Several went further than our own model answers.

Muse and OpenClaw both spotted that Birmingham council tenancies must end on a Sunday, so a 1 December move doesn't fit. Our research had missed that. Gratifyingly, the agents also knew that registering to vote at your new address takes you off the old register, so there's no need to write to your old council.

The mistakes, when they came, were few but telling:

- **Out-of-date rules, delivered with confidence.** Since 6 November 2025, PIP no longer moves across automatically when you move to Scotland. You have to apply for the Scottish replacement, Adult Disability Payment, and your PIP stops 13 weeks after you move. One agent barely searched, said it had "verified" this, and gave the old rule anyway. Anyone following that advice would lose their disability payments.

- **The same agent, opposite answers.** Updating your car's log book doesn't update your car tax direct debit; DVLA says you have to tell them separately. Muse got this right for Hackney and wrong for Leeds a few minutes later, and OpenClaw did the same.
- **Second-hand sources.** The cheapest agent got the answers right, but more than half its links pointed at commercial moving checklist sites rather than the official pages. That's fine until one of those sites is wrong.

As for cost, a thorough search took about two minutes and $0.91 per move. The cheapest set-up did it in 36 seconds for about $0.10.

Companies and charities: the same story, in the small print

Again, the agents found nearly everything: between 92% and 100% of what really mattered, depending on the scenario. They knew to tell the car insurer before moving, to take meter readings on the day, and to book broadband moves early.

Their mistakes were all about details that live deep in a company's help pages or terms:

What the agent said	What the organisation says	Why it matters
Update your Direct Line home insurance online	Moving cover to a new property means phoning them	You think you're insured at the new home when you might not be
National Trust membership doesn't get you into National Trust for Scotland places	Members can visit them too	You might pay to join something you don't need
O2 will want proof of your new address	Only for overseas, PO Box or armed forces addresses	Wasted time digging out documents
Three days' notice is fine for Virgin Media	It asks for at least 30 days	Your broadband may not be ready when you arrive
Your Gift Aid declaration is invalid unless the address is current	A change of address doesn't invalidate it	Needless worry, though you should still tell the charity

None of these are dramatic on their own. But this is precisely the sort of thing a person trusts their agent to get right, and wouldn't think to check. That said, the more expensive the model, the fewer the mistakes it made. And they keep getting better and better.

So what does it mean?

A directory that simply lists who to tell and what to tell them would be politely ignored, and rightly so. Search already does that well enough, cheaply and quickly. The Ai agents might have said they’d like a trusted change of address directory in the user research interviews I did with them, but in practice these tests suggest they probably wouldn’t use it.