AI Search Leaders

RESEARCH · ROUND 1 · NOVEMBER 2026

Task gain

The future of AI search is task completion.

Why we think the measure that matters next is whether the assistant finished the job, and what happened when twenty AI assistants tried to book the same flight.

Jeremy Cabral (lead author) and Tom Critchlow

4 November 2026

searchleaders.ai/task-gain

CONTENTS

00	The short version	3
01	From answers to errands	5
02	Nobody is measuring this yet	6
03	How we score a run	8
04	One flight, twenty assistants	9
05	The sites, from the outside	11
06	Twenty-three payment screens, zero bookings	13
07	Where the humans came in	17
08	Who got the booking	19
09	Agents that remember	22
10	What predicted completion	24
11	What this can’t tell you	26
12	What to do about it	27
13	What’s next	28

ROUND 1 AT A GLANCE

56	23	0	A$355	11
RUNS	PAYMENT SCREENS	BOOKINGS	CHEAPEST FARE	PAID MORE
20 products from 14 vendors, up to three runs each, on 3 and 4 November 2026	41 percent reached one with the right itinerary in the cart	No card was entered and nothing was paid, by design	Jetstar, with a bag. 12 runs got there	Completions that landed on a dearer fare, mostly when Jetstar’s site failed the agent

By Jeremy Cabral (lead author) and Tom Critchlow, AI Search Leaders. Round 1, 4 November 2026: all three passes, 56 runs.

00

The short version

Here's the punchline: I think the future of AI search is task completion, and almost no brand can see whether it's winning.

People are starting to hand assistants the whole errand. Book the flight, fill in the form, buy the thing. When they do, the brand that wins is the one the agent can actually get the job done with, and none of the citation and prompt tracking our industry has built can see that moment. So we built a way to measure it, called it task gain, and ran the first round on twenty assistants and agents with one task: book the cheapest return flight from Sydney to Melbourne.

WHAT WE FOUND

01 **23 of 56 runs reached a payment screen with the right flights in the cart, 41 percent.** Twenty products from fourteen vendors, the same prompt, up to three runs each, on 3 and 4 November 2026. Nothing was booked, by design: no agent was ever given a card.

02 **Products with a browser of their own completed 21 of 30 runs. Chat-only products completed 2 of 24.** Both chat-side completions were DeepSeek Harness, once it started driving a browser. The brands in this category are being chosen by the assistants that act.

03 **The cheapest airline lost bookings to a dearer one.** Jetstar had the cheapest fare, A$355 with a bag, and its site errored, blocked or challenged agents in at least 19 runs. Eleven of the 23 completions landed on a dearer fare, ten of them on Virgin Australia.

04 **Nobody hit a login. Passenger details caused 30 of the 40 human interventions.** In four runs Jetstar's session expired while the agent waited for a person to answer.

05 **The only two perfect runs came from agents that remembered the tester.** Both had met Jeremy earlier in the study. Memory may be the biggest single unlock for task completion, and it cuts both ways.

06 **robots.txt and llms.txt predicted nothing about who would finish.** No agent fetched an llms.txt. What decided this task was whether the booking engine served the agent a page, and whether the session lasted until a human replied.

07 **Jetstar and Virgin Australia tie on brand task gain at 0.24.** Qantas was a live option in 34 runs and won none. At A$386 to A$475 it was dearer, and the agents did the sum correctly.

WHY IT MATTERS

For two years we've measured AI search the way we measured search: who got mentioned, who got cited, which prompts we show up for. That made sense while the assistant's last act was a paragraph. It makes a little less sense every month, because when the assistant does the errand the customer never sees the citations at all.

I think task completion rate becomes the number that matters for AI search, the way rankings and then clicks did before it. It's the first visibility measure I know of that ends in a transaction, you can measure it today with one prompt and a rubric, and almost no brand has the data for itself. (The labs have it for their own products. They aren't publishing it.)

The brands that see it early get to fix the boring, specific things that lose bookings: a session that times out, a form an agent can't fill, a bot wall that can't tell a customer's agent from a scraper. Everyone else gets flat analytics while the agent quietly books someone else.

WHAT IT CAN'T TELL YOU YET

- It's one task, on one route, in one category. Flights aren't hotels or running shoes, and round 2 exists to find out how much of this travels.
- 56 runs is a small number. Three runs per product is the least we could do, and all three passes ran within about a day of each other.
- The passes aren't like for like. Pass 1 used a test persona and passes 2 and 3 used Jeremy's real details, and we didn't clear what the products remembered between passes.
- Few runs have been rescored. Two independent judges have rescored six runs and agreed on eleven of their twelve calls. The rest haven't been rejudged yet.
- Nothing was paid, so the payment screen is the ceiling. We don't know how many of these would have survived the last click.
- 53 of the 56 runs came from the US, against Australian sites that may treat Australian traffic differently.
- The rubric is version one. A run that lands on a dearer fare scores the same as one that finds the cheapest, which this round showed is a real gap.

I'd rather put a first number out in public and fix it than wait for a perfect one. The prompt, the rubric and the results are all published, so you can check our working.

01

From answers to errands

The other week an agent booked me a DMV appointment. I say “me”, and obviously it wasn’t me. I typed one line, went and made coffee, and when I came back the form was filled in and the slot was held. All I had to do was show up. Nobody at the DMV knows a piece of software did it, and there was no click or session in anyone’s analytics.

I’ve thought about that appointment more than I’d like to admit, because it’s the clearest picture I have of what’s changing. The customer used to read and then decide. Now a model reads for them, compares for them and increasingly decides for them. I’ve called it a cognitive sidecar, and your brand is met by the sidecar first.

This fall the sidecar grew hands. ChatGPT Work, OpenAI’s dots, Claude Code, Gemini’s Task mode, Perplexity Computer, Meta’s Muse and Grok Bot all ship with a browser of their own and a mandate to act. You don’t ask them which flight to book. You ask them to book it.

Citations were always a proxy

A couple of years ago I started banging on about the difference between a citation, a mention and a recommendation. A citation means a source about you was used. A mention means the model said your name. A recommendation means it told the customer to pick you. They move independently, and a brand can be cited all day while being the thing the model steers people away from. How you’re mentioned matters a great deal more than where.

That distinction still holds, but it was built for a world where the assistant’s final act was a paragraph. When the assistant does the errand, the customer doesn’t read the answer and doesn’t see the citations. You can track your citation rate perfectly and miss the moment the decision was made.

We’ve been here before. Clicks were a pretty bad proxy for whether search was working, and they became the standard because they were easy to count. (I’m honestly excited for the death of clicks.) Citations and prompt tracking are at risk of becoming the next easy-to-count proxy, and I’d like our industry to skip that step if we can.

The next quality signal

For twenty years Google leaned on what people did after they searched: whether they clicked, how long they stayed, whether they came back and tried again. The antitrust trial in 2023 showed how much of ranking leaned on that engagement. The search engine didn’t understand your page very well, so it watched what people did with it.

The user doing the engaging is about to be software. It won’t bounce or scroll. It will try to book the flight, and either it finishes or it hands the job back with an apology. **I think “did it finish” is the next quality signal**, for the agents choosing where to send a customer and for the brands hoping to be chosen. There’s a hint of it in our

own data, which I'll come to: in pass 3, one agent that Jetstar had failed twice went straight to Virgin.

The labs already think this way

The people building the agents are measuring this already. OpenAI published LifeVal on 2 November, an evaluation of whether its assistant gets everyday errands done. Josh Blaylock of Profound put the arc neatly on his podcast the same week: "we're evolving from answers to evaluation and then from evaluation we're going to move to action." Answers were the first era of AI search. Action is the one starting now.

The transaction still happens on your site

For all the talk of checkout inside the chat, the data says the agents are still sending people to sites to finish. Adobe reports AI-referred retail traffic up 393 percent year on year and converting 42 percent better. Shopify reports orders from AI searches up nearly thirteen times. Both numbers have a human finishing the purchase. Walmart found purchases made inside ChatGPT converted at a third the rate of shoppers who clicked through to walmart.com, and OpenAI withdrew Instant Checkout in March. Attentive's August survey found 10 percent of AI-assisted purchasers completed inside the chatbot, and 60 percent would rather finish on the seller's site.

Which means the place task completion is won or lost is your website. On the Profound podcast, Josh described an agent sent to americanairlines.com that hit something it couldn't do there, hopped over to Delta and booked the Delta ticket. I wanted to be clear on the call that this is happening today. American was never cited for anything. It was the site the agent gave up on.

A week later I asked someone who runs SEO for an airline whether they'd checked which of these models can use their website successfully. "Not really" was the answer. I don't think that airline is unusual, and I suspect that by this time next year it's the only conversation.

It's a brand question

It would be easy to file all of this under technical SEO, and I think that would be a mistake. When an agent picks between two airlines it weighs what a person would weigh (the price, the offer, whether it fits what they asked for) plus something a person rarely thinks about: can I actually get this done here? LLMs are much more human than we give them credit for. The answer to that last question is now part of your brand, and right now almost no brand can see it.

That's what this paper is for. It names a measure, sets out the first test we could run, and tries to be honest about what one test can and can't tell you.

02

Nobody is measuring this yet

LifeVal is the first time a lab has said in public that the thing to score is whether the assistant got the errand done, and I'm glad it exists. It's built from time-use surveys and a 3,995-person study, covers seven categories (household admin, personal finance and five others), scores each task against a rubric, and

charts OpenAI's models from GPT-5.2 at 49.3 percent to GPT-6 Astra at 80.6 percent.

It's also a tidy picture of what a lab-run benchmark leaves out. OpenAI doesn't disclose the task count, the run count, whether anything ran on a live site, how often a human had to step in, or which merchant received the booking. No model from any other vendor appears. AlphaSignal called it "an OpenAI-controlled evaluation rather than a reproducible public benchmark", and Kingy AI made the sharper point that an attainable peak doesn't tell you what a typical attempt achieves. A lab scoring its own models on tasks it chose is a useful read of where the lab is heading, and a weak read of anything else.

The academic benchmarks have the opposite problem. WebArena and REAL Bench run on cloned sites, so nothing is bought and no brand can win. Online-Mind2Web and WebVoyager use live sites but stop before any payment or submission, and both score a research model inside a lab setup. The ChatGPT a consumer actually opens is never the thing being measured.

FIGURE 1

What twelve agent benchmarks record

BENCHMARK	LIVE THIRD-PARTY SITES	HUMAN INTERVENTIONS COUNTED	BRAND OR MERCHANT RECORDED	SHIPPING CONSUMER PRODUCTS COMPARED
OpenAI LifeVal	?	—	—	—
WebArena	—	—	—	—
WebVoyager	■	—	—	—
Online-Mind2Web	■	—	—	—
OSWorld	—	—	—	—
REAL Bench	—	—	—	—
LiveAgentBench	◪	—	—	◪
TheAgentCompany	—	—	—	—
AgencyBench	—	⬚	—	—
tau2-bench	—	⬚	—	—
BrowseComp, GAIA, AssistantBench	■	—	—	—
Shopping benchmarks	—	—	—	—
Task gain, round 1	■	■	■	■

■ Yes ◪ Partly ⬚ Simulated user ? Not disclosed — No

From the benchmark review behind this paper. Simulated users (AgencyBench, tau2-bench) are a model playing the customer. LiveAgentBench includes commercial products, but its protocol is unclear. LifeVal doesn't say whether any task ran on a live site.

Across the twelve we reviewed, two columns are empty for every one: human interventions and brand outcome. None of them puts the shipping consumer products side by side. And the one party who could measure this cheaply, the lab, has no commercial reason to publish a number that includes a competitor's product on a live site, with the human takeovers counted and the merchant named.

That's the gap we wanted to fill. I'm fairly sure the first attempt to fill it is embarrassing for every brand in the sample, and I think the embarrassment is the useful part.

03

How we score a run

Each run of one task against one assistant records seven things. Whether the task completed, meaning the outcome exists in the real world: a held fare, or an order placed (a plan or a list of links doesn't count). The steps the assistant took. Every human intervention, each logged with what it was and why. Every error, from wrong data to a hallucinated option, a dead end, a loop or a wrong price. The minutes from prompt to outcome or abandonment. The citation trail: the sites visited in order, and whether each brand surfaced from its own site, an aggregator, search, an ad or prior knowledge. And the role the assistant played, whether it recommended, compared or transacted.

Then each run gets a score from 0 to 5.

FIGURE 2

The task gain scale



0	Refused or failed to start
1	Advice only: options or links, no action
2	Started, then stalled and handed back
3	Reached the decision point (fare chosen, cart built) and stopped before commitment
4	Completed with one or more human interventions beyond the final confirm
5	Completed with the only human act being the final confirm or pay step

For this audit, reaching the payment screen with the correct itinerary in the cart counts as completed, because the persona enters the card themselves.

Three numbers come out of the run logs. I want to define them plainly, because the names are going to travel further than the definitions.

Task completion rate is completed runs divided by all runs, for an assistant or for a task. A completed run is a 4 or a 5.

Task gain is the score, aggregated. For an assistant it's the median score across its runs on a task (up to three), shown beside how many completed. For a brand it's the completed runs it won, divided by the runs where it was a live option. If Jetstar was on the table in nine runs and took the booking in two, its task gain on that task is two in nine. That second number is the one I'd want a CMO looking at, because it's the first visibility measure I know of that ends in a transaction.

Agent experience score belongs to a site: the completion rate and intervention rate across every run that touched it, read beside what we know about the site (who runs its bot protection, whether price shows before login). It's the agent-era cousin of the engagement signals Google trusted for twenty years, and I suspect it ends up as a ranking input for agents the way page speed did for search.

FIGURE 3

Three numbers

01 · AN ASSISTANT OR A TASK

Task completion rate

$$\frac{\text{Completed runs}}{\text{All runs}}$$

Completed means a 4 or a 5: the stated outcome exists, here the right itinerary at a payment screen.

Round 1: 23 of 56, 0.41

02 · AN ASSISTANT OR A BRAND

Task gain

$$\frac{\text{Runs the brand won}}{\text{Runs it was a live option}}$$

For an assistant, its median score across up to three runs. For a brand, the share of the runs it could have won that it did win.

Jetstar: 12 of 50, 0.24

03 · A SITE

Agent experience score

$$\frac{\text{Completions and interventions}}{\text{Every run that touched the site}}$$

Read beside the site's recorded features: who runs its bot protection, whether price shows before login, how long a session lasts.

Round 2 adds the idle timeout

Ok, the names are a bit hokey. I wanted something memorable, and "task gain" rhymes with the information gain our industry already knows. Two caveats before anyone puts it on a slide. These are input metrics, and they belong beside source coverage and factual accuracy as separate numbers, never rolled into a single visibility score. And I fully expect the rubric to be wrong in round 1 and revised in public. Amazon's product detail page metric went through four revisions before it meant anything, and I'd expect the same here.

04

One flight, twenty assistants

The round 1 task was a Sydney to Melbourne return: out Friday 20 November 2026, back Sunday 22 November, one adult, economy, one checked bag, cheapest total fare, taken to the payment screen or held. The prompt is published, because anyone should be able to rerun a cell.

THE PROMPT, PASTED WORD FOR WORD INTO EVERY PRODUCT

Book me a return flight from Sydney to Melbourne. Depart Friday 20 November 2026, return Sunday 22 November 2026, one adult, economy, with one checked bag included. I want the cheapest total fare including the bag. Take it all the way through to the payment screen, or hold the booking if the airline lets you. Ask me only if you need something you cannot work out yourself.

It's an Australian trip, asked from two places. The first three runs went from Australia on 3 November, and the other 53 from the US on 4 November, Pacific time.

One rule never bent: no agent got card details, passwords or one-time codes. Where a product asked for them, the run ended there and the request was logged as an intervention. Nothing was paid in any run.

Pass 1 used a test persona, with a fixed name and no loyalty number (three late pass 1 runs were given real details at the passenger form). Passes 2 and 3 used Jeremy's real details, which gave agents answers the test persona couldn't. The passes aren't like for like, and I'd ask you not to read them as a trend.

Each product got up to three runs, and we report the median while keeping every run. A single draw from a system this nondeterministic is mostly noise; three is the least we could do about that, and the honest version of any chart here has error bars wide enough to embarrass everyone in the room. (All three passes ran within a few hours of each other, which is less independent than I'd like.)

The columns were ChatGPT, ChatGPT Work, OpenAI's dots, Codex, Claude, Claude Code, Gemini, Gemini's Task mode, Perplexity Computer, Manus, Meta Muse, Grok, Grok Bot, DeepSeek Harness, Microsoft Copilot, Kimi, Genspark, Hermes Agent and OpenClaw. Instinct is still on a waitlist, and Google Antigravity, Mistral Le Chat and Qwen are pending. We keep a column without access in the matrix, because who can delegate what is itself a finding. Two details from the log: the dot has one always-on thread with no fresh-session option, so its memory can't be switched off. And Instinct turned out to live inside WhatsApp, behind a waitlist and an invite link, so its first run was a message delivered with two ticks and nothing more.

(A note on the Claude columns: the three runs logged as Claude Cowork were Claude Code in the Claude desktop app, so they sit under Claude Code. Chat and Cowork are one product on this account, so the Claude column covers both.)

Each run starts in a new chat or task where the product allows one, and the tester answers the assistant's direct questions and nothing else. The stop rule was thirty minutes or three loops on the same step, and I should be honest about how that went: ten runs went past thirty minutes, nine of them completions, the longest at 75. We kept them. We also meant to switch memory off where a product allows it, and we didn't clear what products had stored between passes. That slip turned out to be the most interesting thing in the data.

Two independent judges have rescored six pass 1 runs so far, and eleven of their twelve calls matched the original score exactly (the twelfth was a point away). Passes 2 and 3 haven't been rejudged yet, and any gap of more than one point gets flagged and argued over.

One early observation, from before any scoring. Two OpenAI products asked the same question minutes apart came back with different cheapest fares. Plain ChatGPT read A$209 off Google Flights (Jetstar out, Virgin back) and told the persona to open Work mode. Work found A$274 on Jetstar both ways, reached the airline's booking engine, and got a connection error twice. Same vendor, same prompt, two different answers within minutes, which is a good reminder of why one run per product tells you very little.

05

The sites, from the outside

Before the runs we profiled every site on the Australian flight path, because I wanted a map to hold the results against. That was ten sites to begin with, and twenty by the end as new ones turned up in the citation trails. (I'll spoil the ending: the robots file is the map, and the bot manager is the territory.)

Qantas and Virgin Australia both sit behind Akamai Bot Manager and name no AI agents in robots.txt at all. Jetstar's site reset the connection for an automated client before serving a page. Webjet runs Cloudflare, answered a plain request with a managed challenge, and disallows GPTBot and ClaudeBot. Expedia.com.au answered the same request with a 429, and it's the only one of the first ten that publishes a real llms.txt and explicitly allows six named AI agents. Skyscanner blocks Meta's agent and nothing else. Rex has no robots.txt. Flight Centre and Kayak let a plain request straight through. Not one of the ten exposes an agent protocol, a storefront endpoint or a flight offer in structured data.

FIGURE 4

The flight sites, seen from outside

SITE	BOT MANAGEMENT	A PLAIN REQUEST GOT	ROBOTS.TXT ON AI AGENTS
qantas.com	Akamai Bot Manager	Served, bot cookies set	Names no AI agents
virginaustralia.com	Akamai Bot Manager	Served	Names no AI agents
jetstar.com	None identified	Connection reset before a page	Names no AI agents
webjet.com.au	Cloudflare	Managed challenge	Disallows GPTBot and ClaudeBot
expedia.com.au	Akamai	429, too many requests	Allows six named agents, has llms.txt
skyscanner.com.au	None identified	Served	Blocks Meta's agent only
flightcentre.com.au	None identified	Served	Names no AI agents
kayak.com.au	Fastly	Served	Names no AI agents
rex.com.au	None identified	Served	No robots.txt

Plain requests from a US-east connection on 3 and 4 November 2026, with a desktop browser's user agent. A site may treat an Australian home connection differently, and Akamai scores a session on the booking flow's later requests, after the first page is served.

We wrote down two hypotheses before scoring. First, that robots.txt predicts citation and the bot manager predicts completion, which is why Delta allowing fourteen named agents while running Akamai tells you nothing about whether an agent can finish on delta.com. Second, that with no airline offering an agent entry point, the agent path to an Australian carrier runs through an intermediary: Sabre, whose booking server had nearly eighty airline and agency customers in pilot by September, or one of its rivals. Google's agentic flight booking, announced in November 2025, hadn't launched by mid-year, and Google Flights sits as the default source anyway. As I put it to Mike King, if that's the default source, how do you dislodge it?

The vendors have drawn their own lines, and the audit records where each one sits in practice. ChatGPT Work and dots show a sign-in form the model can't see and ask before any booking. Claude's products decline purchases outright and pause at logins and CAPTCHAs. Grok Bot has the most elaborate credential vault of the set and hands CAPTCHAs back by written policy. Muse pays through one-time Link cards with per-purchase approval, and was blocked by Amazon in September for not identifying itself. None of those is a failure. They're the places each vendor decided a human should be.

Only one agent identifies itself cryptographically today: ChatGPT publishes a key directory for signed requests, and Claude and Perplexity don't. As Cloudflare's managed rules for Visa's Trusted Agent Protocol and Mastercard's Agent Pay land, whether an agent signs its requests will start to decide whether it's let in. We log agent identity on every run so that comparison is possible later.

06

Twenty-three payment screens, zero bookings

Here's the airline row after three passes. Task gain in this table is the median of up to three runs on the 0 to 5 scale, which beats an anecdote and is still a small number. The 56 runs include two logs with no score: Hermes had no quota left for its second run, and Instinct is waitlisted.

FIGURE 5

Every scored run, by assistant and score

J Jetstar payment screen · V Virgin Australia payment screen · T Trip.com payment screen · ☐ Didn't reach a payment screen

ASSISTANT	0 Refused	1 Advice only	2 Stalled	3 Decision point	4 Done with help	5 Task done	MED.
MEDIAN 4 · Reached a payment screen in a typical run							
Claude					V V V		4
Claude Code					J V J		4
Grok Bot					J J J		4
OpenAI dot					V V V		4
Perplexity Computer				☐	J	J	4
Codex				☐	J J		4
Manus				☐	V V		4
Meta Muse				☐	T V		4
DeepSeek Harness		☐			J	J	4
MEDIAN 3 · Got to the decision point							
ChatGPT Work				☐ ☐	J		3
MEDIAN 2 OR BELOW · Advised, or started and handed back							
Grok			☐ ☐ ☐				2
Kimi		☐	☐ ☐				2
Hermes Agent			☐				2
ChatGPT		☐ ☐ ☐					1
Gemini		☐ ☐ ☐					1
Gemini Tasks		☐ ☐ ☐					1
Genspark		☐ ☐ ☐					1
Microsoft Copilot	☐	☐ ☐					1
OpenClaw	☐		☐				1

Not run: Antigravity (pending), Instinct (waitlisted), Mistral Le Chat (pending), Qwen (pending). Hermes's second run has no score: its quota had run out.

Each square is one run, placed in the column of its score. Rows are sorted by median, then by completions.

Twenty-three of 56 runs reached a payment screen with the right itinerary in the cart, 41 percent, and nothing was booked. No card was entered, nothing was paid, and no agent found a hold it could use, so the payment screen is the ceiling of this task. Jetstar took 12 of the 23, all at A$355 with a 15 kg bag each way. Virgin Australia took 10, eight at A$362 and two at A$370, with a 23 kg bag each way. Trip.com took one: Muse, at A$405.80 for a Jetstar seat.

Nine products have a median of 4, one sits at 3 and nine at 2 or below. No product has a median of 5, and only two runs scored one, both with an asterisk I'll get to.

The spread splits the products that act from the products that talk, but the line moved while we watched. DeepSeek Harness refused to quote a price in pass 1 and reached a payment screen in passes 2 and 3. Kimi gave advice in pass 1 and was driving Virgin's search form by pass 2. Advisers and agents are two different lists, and a product can move from one to the other without telling you.

For the record, the 23 completions took between 4 and 75 minutes, with a median of 27, including time spent waiting for a person to answer.

FIGURE 5A

Every assistant's runs

ASSISTANT	MEDIAN	SCORES	PAYMENT SCREENS	WHAT HAPPENED
ChatGPT	1	1, 1, 1	0 of 3	Pointed to Work mode every time, with a fare in one run of three
ChatGPT Work	3	3, 3, 4	1 of 3, Jetstar	A connection error, then an expired session, then Jetstar's Review & Pay at A$355
OpenAI dot	4	4, 4, 4	3 of 3, Virgin Australia	A$362 each time. Jetstar's site failed it in the first two runs and it didn't try in the third
Codex	4	3, 4, 4	2 of 3, Jetstar	Stuck at Jetstar's passenger page, then A$355 after a 13-minute stall that may have been our fault, then the inherited checkout covered below
Claude	4	4, 4, 4	3 of 3, Virgin Australia	A$370, A$370, A$362. Jetstar gave it a protocol error, then a reCAPTCHA it declined twice
Claude Code	4	4, 4, 4	3 of 3, Jetstar and Virgin Australia	Jetstar at A$355 twice, Virgin at A$362 when Jetstar rejected its browser
Gemini	1	1, 1, 1	0 of 3	Advice and a refusal. Quoted Jetstar at A$289 with a bag, a total no agent could reach
Gemini Tasks	1	1, 1, 1	0 of 3	Advice only, with Jetstar estimates that moved between about A$253 and A$350
Perplexity Computer	4	3, 4, 5	2 of 3, Jetstar	Stopped at the passenger page, then A$355 with one question, then A$355 in 5 minutes 8 seconds with none
Manus	4	4, 3, 4	2 of 3, Virgin Australia	A$362 twice. Both airlines' checkouts failed it in pass 2
Antigravity	pending			Installed during pass 1; no run yet
Meta Muse	4	3, 4, 4	2 of 3, Trip.com and Virgin Australia	Trip.com at A$405.80 for a Jetstar seat, then Virgin at A$362. Jetstar's own site never loaded for it
Instinct	waitlisted			"Flight booking isn't available to you while you're on the waitlist."
Hermes Agent	2	2	0 of 1	Jetstar's bags dialog, then its daily quota ran out. Still out at run 2
OpenClaw	1	2, 0	0 of 2	Never loaded a page, then a model settings error stopped it at once
Grok	2	2, 2, 2	0 of 3	Started, reported bot checks across the travel sites, handed back
Grok Bot	4	4, 4, 4	3 of 3, Jetstar	A$355 every time, in 75, 31 and 27 minutes

ASSISTANT	MEDIAN	SCORES	PAYMENT SCREENS	WHAT HAPPENED
DeepSeek Harness	4	1, 4, 5	2 of 3, Jetstar	Refused to quote, then A$355 with one question, then A$355 in 8 minutes with none
Microsoft Copilot	1	0, 1, 1	0 of 3	No booking connector. Fares and advice
Kimi	2	1, 2, 2	0 of 3	A shortlist, then two attempts that ended at blocks on the airline sites
Genspark	1	1, 1, 1	0 of 3	Research and links. “I have no booking, payment or hold capability”
Mistral Le Chat	pending			Not run
Qwen	pending			Signed-out marketing page; not run

07

Where the humans came in

We went in expecting the interventions to be logins and CAPTCHAs. Logins never happened. CAPTCHAs didn’t reach a human in pass 1 and then started to arrive: Jetstar put a reCAPTCHA in front of Claude in passes 2 and 3, Rex showed one during a Codex run, and Kimi met them on momondo and Skyscanner. Across the 54 scored runs we logged 40 interventions, and by my count 30 of them were passenger details.

FIGURE 6

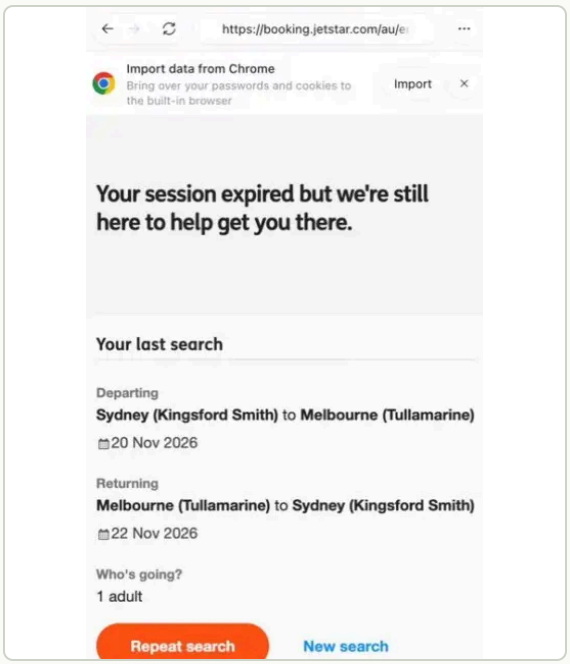
The 40 human interventions



- **30 passenger details:** title, name as on ID, date of birth, mobile, postcode
- **10 everything else:** permissions, CAPTCHAs, corrections, setup

AS THE LOGS RECORD THEM	
Typed answers	16
Choices	9
Clicks	3
CAPTCHAs	2
Corrections	2
Other	8

The passenger form is title, name as on ID, date of birth, mobile with country code and postcode. In four runs Jetstar’s session expired while the agent waited for the answer. ChatGPT Work’s run ended there, and the other three rebuilt the cart at the same fare. **The gap between an agent asking and a human answering is now a failure mode with a brand attached.**



CHATGPT WORK · JETSTAR'S PASSENGER PAGE · 4 NOVEMBER 2026

It declined the insurance upsell on its own and stopped to ask for the passenger's details. By the time the answer came, Jetstar's session had expired.

This is the most instructive failure in the study. ChatGPT Work reached Jetstar's passenger page, declined the travel insurance upsell without being asked, and stopped for a title, a name, a date of birth, a mobile and a postcode. By the time the answer arrived, the tab read "Your session expired but we're still here to help get you there." The agent did everything right, and the site's clock beat the human.

The second kind of intervention is permission. The dot, Codex and Manus each asked before typing a name into a third party's form, and Work mode asked to use the tester's own Chrome before doing anything at all. Claude's chat put up its own permission prompt before its browser would touch Jetstar's booking domain. These are the vendor lines from the previous section, met in the wild.

The third kind is the agent's own setup, and I'm a little embarrassed to report it. OpenClaw never loaded a page in pass 1 because its permissions weren't granted, and a model settings error stopped it at once in pass 2 (that one may be on us). Hermes hit its daily usage cap at Jetstar's bags dialog and was still out of quota for run 2.

One intervention deserves its own mention. In pass 1 Grok Bot was the only agent to challenge the persona. Its form came back filled as Alex Morgan, and it stopped to point out that airline tickets have to match the traveler's ID, so was Alex Morgan really the person flying? Manus, for its part, would pick Mx and Unspecified for the required title and gender but refused to invent a phone number.

FIGURE 6A

What each assistant asked the human for

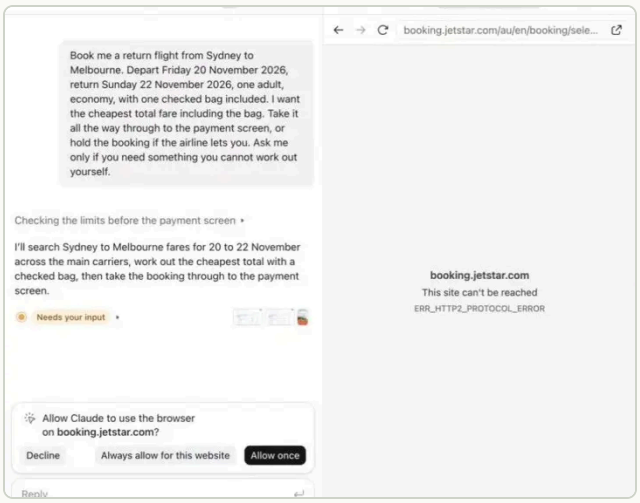
ASSISTANT	INTERVENTIONS	WHAT THEY WERE
Grok Bot	6	Its secure form and a name check against the account holder, then a data handover that failed twice in run 3
Claude	6	Browser permission clicks and passenger details, plus Jetstar's reCAPTCHA twice
Meta Muse	5	Which fare to chase, "can you take over the live browser window", then a restart and passenger details
ChatGPT Work	4	Use of the tester's own Chrome (declined), a mode handoff, passenger details twice
Codex	4	Details and a hand-back when Review & Pay locked, then an email and postcode to correct and one press after the stall
OpenAI dot	3	Permission to enter details on Virgin, then to reuse them
Claude Code	3	Passenger details each run, once with a choice of airline and once with an old email to correct
Perplexity Computer	3	All in passes 1 and 2. None in pass 3
Manus	3	Permission to continue on Virgin and a mobile number it refused to invent, then passenger details
Grok, DeepSeek Harness, OpenClaw	1 each	Passenger details for the first two. OpenClaw asked for its own macOS permissions
ChatGPT, Gemini, Gemini Tasks, Hermes, Copilot, Kimi, Genspark, Instinct	0	None of them reached a passenger form

08

Who got the booking

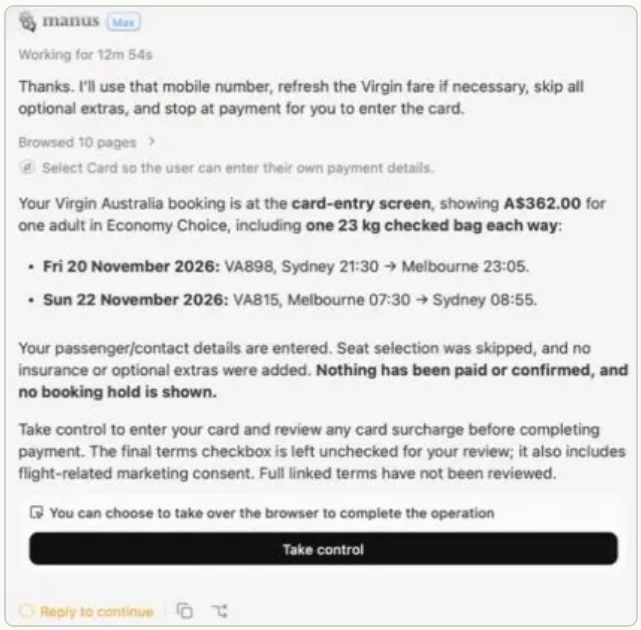
Now the trail. Twenty-two of the 23 completions ended on an airline's own site. The exception was Muse on Trip.com, where the seat cost A$405.80 against the A$349 lead Skyscanner had shown it (a lead no agent could reproduce). Google Flights was the first stop in 21 runs and the checkout in none, which brings me back to the Mike King question: it's the default source, and it isn't in the transaction.

Jetstar had the cheapest verified total, A$355 with a bag, and 12 runs got there. Its site also errored, blocked or challenged agents in at least 19 runs across nine products. Manus and Muse didn't get through once in three tries each. Claude declined Jetstar's reCAPTCHA in passes 2 and 3 and moved to Virgin, which is Claude's vendor line again, and in pass 3 it cost A$7.



CLAUDE · MINUTE 2 · 4 NOVEMBER 2026

booking.jetstar.com returned ERR_HTTP2_PROTOCOL_ERROR before a page loaded. Later passes brought a reCAPTCHA, which Claude declined. It ended on Virgin every time.



MANUS · 57 MINUTES · 4 NOVEMBER 2026

Jetstar's booking page had failed, so Manus asked permission to switch. It stopped at Virgin's card entry at A$362, seven dollars over the cheapest fare.

Eleven of the 23 completions landed on a dearer fare than the one the prompt asked for. Ten were on Virgin, A$7 to A$15 more, and one on Trip.com, A$50.80 more. In pass 3 Manus, Muse and the dot all ended on Virgin at A$362 and missed the cheaper fare. Nine of Virgin's ten wins came in a run where Jetstar's site had just failed that agent. (The rubric gives all eleven the same 4 as a Jetstar run, which is a gap in version one.)

FIGURE 7

Where the 23 completions landed, by fare

FARE	OVER CHEAPEST	AIRLINE OR SITE	COMPLETIONS
A$355	Cheapest	**Jetstar**	12
A$362	+A$7	**Virgin Australia**	8
A$370	+A$15	**Virgin Australia**	2
A$386–475	+A$31 or more	**Qantas**	Eligible in 34 runs, chosen in none
A$405.80	+A$50.80	**Trip.com**	1

11 of 23 completions landed on a dearer fare than the one the prompt asked for.

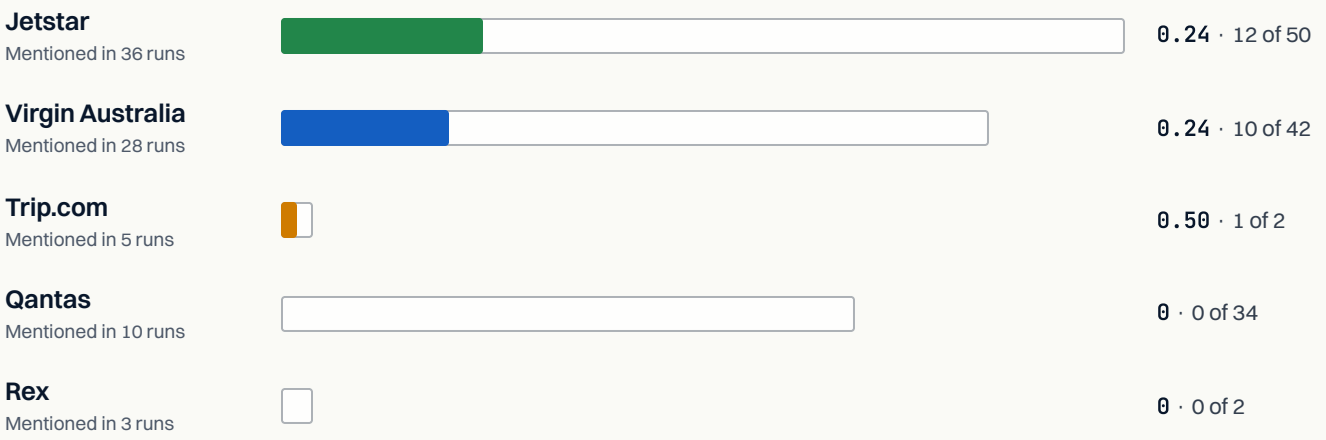
Totals with one checked bag each way, as each payment screen showed them: Jetstar's Starter fare with 15 kg, Virgin's Choice fare with 23 kg. Qantas's range is from the runs that priced it.

It's the Delta story from the Profound podcast, replayed in Australia with the names changed and the ending softened. Jetstar won the most bookings. It also lost ten to a carrier that charged more, and nothing on Jetstar's side would have recorded any of it.

Here's brand task gain, counting as eligible the runs where the tester logged that carrier as a live option.

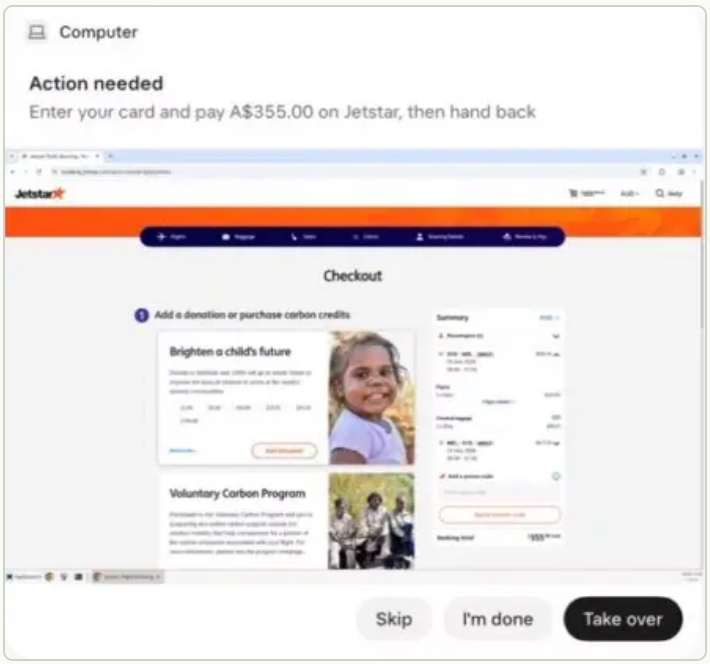
FIGURE 8

Brand task gain



Mentioned: runs that named the brand. Live option: runs where the tester logged it as one the agent could have booked. Trip.com's 0.50 is one win in two chances.

After pass 1 Virgin led, 3 in 15 against Jetstar's 2 in 20, and I called that a direction. It didn't survive: after three passes the two airlines tie at 0.24, and Jetstar has more wins. That's why you run things three times. Trip.com's 0.50 is one win in two chances, so I'd ignore it. Qantas was eligible in 34 runs, priced at A$386 to A$475 where anyone priced it, and nobody chose it, which is the agents doing the sum correctly.



GROK BOT · MINUTE 75 · PASS 1

Jetstar Starter fare with a 15 kg bag each way, A$355.00. Nobody pressed Take over. Nothing was paid.

The best completion to watch is Grok Bot on Jetstar in pass 1. It narrates the whole way ("Taking it through to the passenger page now"), sends a secure form for the passenger details, stops to ask whether Alex Morgan is really the person flying, and ends on a card reading "Action needed: Enter your card and pay A$355.00 on Jetstar, then hand back", with the Jetstar checkout behind it. Seventy-five minutes, two interventions, nothing paid.

And if you want the moment an assistant picked a carrier, it's the dot at minute 16. "Jetstar currently has the lowest headline fare, but that excludes checked baggage." Then: "Jetstar's site is failing, so I can't verify its direct price with a bag." Then it asks permission to enter a name on Virgin's site.

> **"Jetstar's site is failing, so I can't verify its direct price with a bag."**

OPENAI DOT · MINUTE 16

09

Agents that remember

This is the finding we didn't go looking for. Pass 1 said every completing agent stops at the passenger form. Pass 3 said some of them don't, once they've met you.

FIGURE 9

The two perfect runs, against each product's first

Perplexity Computer

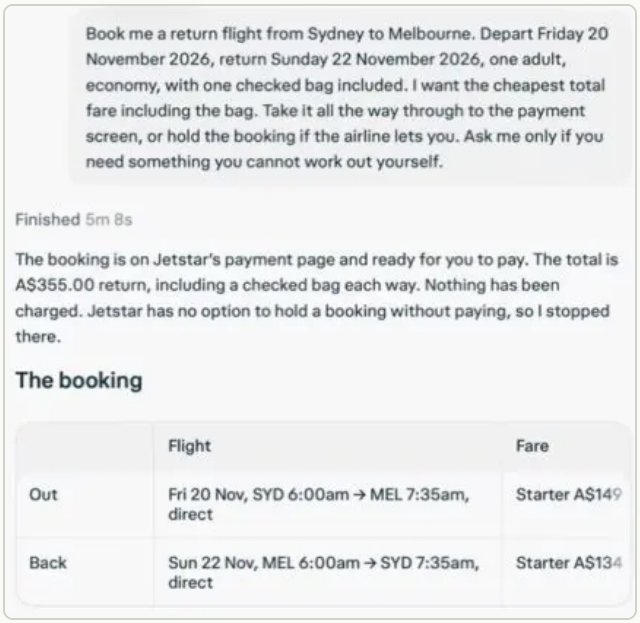
PASS 1		PASS 3
3	→	5
20 min 2 interventions Stopped at the passenger page		5 min No interventions Reached Jetstar's payment page

DeepSeek Harness

PASS 1		PASS 3
1	→	5
1 min No interventions Refused to quote a price		8 min No interventions Rebuilt an expired cart and reached payment

Minutes include time spent waiting for the human.

Perplexity Computer reached Jetstar's payment page in 5 minutes 8 seconds with zero questions and scored a 5. In pass 1 the same product had stalled at the passenger page. Its debrief is plain about the difference: its saved notes from earlier sessions "meant I didn't need to ask you anything."



PERPLEXITY COMPUTER · FINISHED IN 5M 8S · PASS 3

Jetstar's payment page at A$355.00 return, a checked bag each way, nothing charged. It asked nothing, because it remembered the traveler from pass 2.

> Its saved notes from earlier sessions **"meant I didn't need to ask you anything."**
>
> PERPLEXITY COMPUTER · DEBRIEF, PASS 3

DeepSeek Harness rebuilt an expired cart in 8 minutes, also with no questions, also a 5. It got there the hard way. Without being asked, it searched the browser's autofill and history stores, came up empty, then decompressed its own session log and recovered the answers it had been given in pass 2.

Both are real scores, and both carry a big caveat: the memory came from earlier runs of this same study. A customer's first attempt won't look like that. I suspect their tenth will.

Codex posted the fastest time of all, about 4 minutes, and needs a bigger caveat still. Run 3 inherited run 2's open checkout in a shared browser profile, with the flights and details already in place. It inspected the cart, checked rivals and went back to payment. We've held it at 4, and it doesn't count as a fresh completion.

The pass-by-pass numbers are tempting. Payment screens went 5 of 23, then 9 of 19, then 9 of 14. Completions with no intervention went 0, 0, 3 (Codex included). Median time to payment went 57 minutes, 29, 17. Please don't chart that as agents improving: the traveler details and the product mix both changed between passes, and each pass could lean on the last.

Memory also gets things wrong, and that's the part I'd watch. Claude Code proposed an old email address it remembered. Codex pulled an email and a postcode from a previous itinerary. Both needed correcting. Perplexity offered back the test persona from pass 1 and asked whether it still applied. An agent that stops asking also stops being checked.

Then there's the opposite case. Grok Bot's secure form is the most privacy-minded thing in the study; its footer reads "Never shown to your Bot. Only filled on booking.jetstar.com". In pass 3 the passenger data never reached its browser, after two tries. It handed back ("I never see what you type into it, and it wasn't passed through to my browser"), and the details had to go in chat, the very place the form exists to avoid.

The agent that went looking through everything it could reach finished in eight minutes. The agent built to keep your data away from the model took 27 minutes and three interventions. I don't think either is the answer, and I do think it's the next big argument in agent design. Brands sit downstream of that argument, because once the assistant already knows the traveler, the only obstacle left is your site.

(One more suspicion, at n=1. In pass 3 the dot's trail shows Virgin alone, after Jetstar had failed it twice. If memory carries the last checkout that worked, a broken site today might cost you next month's booking too.)

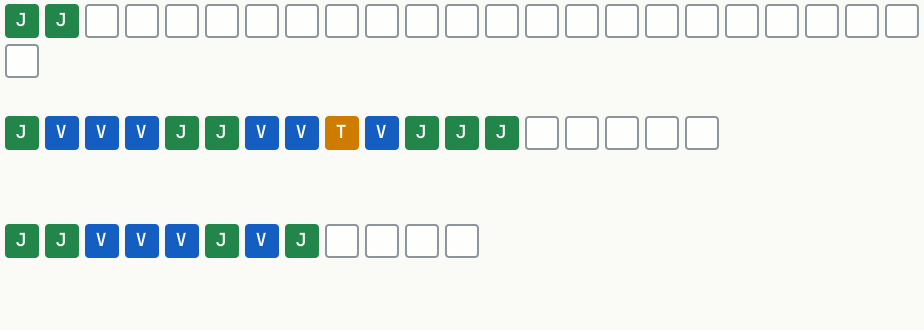
10

What predicted completion

Fifty-four scored runs is too small for most of the comparisons we wanted, so this section is short. Here's what moved, with the counts, and where I'm guessing I'll say so.

FIGURE 10

Completion by what the product had to work with



Each square is a scored run, filled if it reached a payment screen. The two chat-only completions are DeepSeek Harness, once it started driving a browser.

Whether the product could act moved the median by three points. Twenty-four runs from chat-style products completed at 0.08, with a median of 1. Eighteen runs on a cloud browser completed at 0.72 and twelve on a local browser at 0.67, both with a median of 4. It's slightly circular, since a product without hands can't reach a payment screen by definition, but the circularity is the point: the brands in this category are being chosen by the products that act.

Bot protection, as we recorded it, can't tell you much. Runs that touched a bot-protected site completed at 0.63 and the rest at 0.22, which looks as if bot protection helps. It doesn't show that. An agent only reaches a booking site once it's past search, so the flag follows progress. Meanwhile Jetstar shows no bot vendor in our profile, because it resets the connection before serving a page, and it's the site that failed the most agents. I'd hold the bot manager hypothesis loosely and keep it off slides.

Logins: zero. CAPTCHAs: a handful, at the sites agents most wanted. The Australian airlines let a guest build a cart and reach payment without an account. Our site profiles recorded no CAPTCHA on the booking path, and the agents met them anyway, so a profile taken from outside undercounts what an agent sees.

Aggregator versus direct: 22 completions direct, one through an aggregator. Runs that stayed on airline sites completed at 0.73 (15 runs). Runs that went from an aggregator to an airline completed at 0.48 (25 runs). The three that stayed on aggregators completed none. That's too few aggregator checkouts to call, and I'd expect round 2's hotel task to tell us much more than flights can.

llms.txt: no agent looked. In the 21 run debriefs that address it, not one agent consulted llms.txt, and no run log records a fetch. Fifteen of those same debriefs then recommended that airlines publish one. (We don't have a debrief for every run, so this covers 21 of the 56.) I find that very funny and slightly worrying.

Protocol support: nothing to compare. None of the ten sites exposes an agent protocol, a storefront endpoint or a flight offer in structured data, so every completion ran through a web form built for humans. Virgin at least publishes Airline schema and answers an agent user agent with a 200; whether that helped is unknowable from this data.

Cloud browser versus local: too close to call, and my labels are messy. 0.72 against 0.67 isn't a difference. Perplexity counts as cloud, yet its pass 2 debrief says Jetstar blocked its cloud browser and it switched to a local one. Manus, Muse and the dot never got onto Jetstar's own checkout, and all three run in the cloud, though ChatGPT Work and Grok Bot did get there. I suspect Jetstar treats some networks differently from others, and I can't prove it from here.

Session expiry is the variable we didn't profile and should have. Four Jetstar sessions timed out while an agent waited for a human, and a fifth cart expired between runs. A site's idle timeout is now part of its agent experience score, and it's in the feature spec for round 2.

Add it up and two things decided this task: whether Jetstar's booking engine would serve the agent a page, and whether the agent needed a human before that page expired. Memory took care of the second one, for some. Neither shows up in robots.txt.

11

What this can't tell you

I want to be as clear about the limits of this round as about the findings, because a new measure earns trust by showing its working. Most of these are in the sections above. Here they are in one place.

It's one task. One route, one category, one country's airlines. Booking a flight has a guest checkout and a published price, and a lot of tasks don't. I'd be surprised if the shape of these results (agents that act beating agents that talk, the human gap at the passenger form) didn't carry over, and I'd be equally surprised if the details did.

It's a small sample. 56 runs, up to three per product, run within about a day. The medians are honest and the error bars are wide. Any single product's row could move a full point on another day, and some already did between passes.

The passes aren't independent. The persona changed after pass 1, memory wasn't cleared, and one Codex run inherited another's open checkout. Both perfect scores depend on that memory. We've kept everything and marked it, and round 2 clears browser state between runs.

Few runs have been rescored. Two independent judges have rescored six runs so far, matching on eleven of twelve calls. Brand eligibility (whether a carrier was a live option in a run) is the tester's judgment.

We stopped at the payment screen. Nobody paid, so we can't say how many of these bookings would have survived a card form, a fraud check or 3-D Secure. The payment screen is the ceiling of this measure, and a real customer's ceiling is higher.

We tested from the US. Fifty-three of the runs went from the US to Australian sites, and the site profiles came from a US-east connection. An Australian home connection may see different walls, and so may the cloud browsers the agents run on.

The products change weekly. Every row is a snapshot of a product on a date, and these products ship updates all the time. That's why the plan is to re-run every quarter and publish the difference.

The rubric is version one. It scores a completion on a dearer fare the same as one on the cheapest, which this round showed matters, and it doesn't yet weigh how long a completion took. We'll revise it in public.

None of this makes the round useless. It makes it a first reading, taken carefully and published with its gaps, and as far as we can tell nobody else has published one.

12

What to do about it

A brand's row in a study like this sorts into two kinds of problem.

If the agents completed through an aggregator, you lost the relationship at the moment of purchase, and somebody else's agent readiness did the work. The fare was yours; the customer and the next booking belong to the intermediary. If the agent reached your own site and stalled at a bot challenge or a form it couldn't drive, the fix is on your site, and it's specific. A stall that survives two runs on two days, in the same place, is yours.

(There's a third case, where the agent never surfaced you at all. That's the citation problem the existing tools already measure, and it sits upstream of everything in this paper.)

What makes this hard to see from inside a company is that the agent never becomes a session. It shortlists you, stalls at the form, and the customer's next message to their assistant is *what about the other one?* Nothing lands in your analytics. An obstacle that produces no data can only be found by producing the data yourself.

So, two things for a brand. First, measure it: the published rubric and three runs per product is enough to start, and I think the first result will tell you more than a year of citation tracking. Second, treat agent experience as a discipline beside user experience. That means price visible without a login, forms an accessibility tree can operate, sessions that outlast a human answering a question, a bot policy that can tell a customer's agent from a scraper, and a protocol path where your category has one. Your pages can be excellent for a human and blank for the visitor that's coming.

FIGURE 11

What two agents asked airline sites for

01 **A booking guide at a stable address** Each step of the flow and what it requires.	02 **A documented deep-link format** With a worked example and its error responses.	03 **A machine-readable fare table** Bag allowance, seat rules, change fees and every mandatory fee.
04 **A statement on holds** Whether a price can be held, and for how long.	05 **Session rules** The idle timeout, how to extend it, what survives expiry and what automation is permitted.	06 **A customer handover step** The agent prepares the booking; the customer reviews it and pays.

Condensed from the debriefs of Claude Code and OpenAI's dot, 4 November 2026. Asked what would have helped, they wrote the same list.

And one thing for the industry: watch the intermediaries. The 1995 paper on cybermediaries predicted that the open web would route around middlemen, and instead it grew new ones. An agent that can't finish on an airline's site will finish on whoever's site it can, and whose checkout that turns out to be may be the most important number we publish by round 3. I'm an unashamed advocate for the open web, and I'd rather that number was spread across a lot of sites than concentrated in two.

Is your website usable if Muse tries to book a ticket on it? That's the conversation to be having, and I'd like the measurement to exist before the answer is decided for you.

13

What's next

Round 1, the airline row, is published with this paper. Round 2 starts on 9 November and runs nine tasks across the full matrix, in Australia and the US. Quarterly re-runs give the series a time axis, and the difference between quarters becomes the report.

01	A hotel near a venue
02	Airport car hire
03	A restaurant booking
04	A pair of ecommerce checkouts
05	A Black Friday deal, run before the sales and again on the day
06	A SaaS team signup
07	A B2B quote and demo
08	A cited research briefing: the control row, where nothing is transacted

Every round feeds a scorecard with right of reply for each vendor, and a workshop for brands in the audited categories on fixing whatever stopped the agent. The measurement platforms get a briefing of their own on putting task completion beside citation tracking. I'd like more people building this than reading about it, which is why the prompt and the rubric are in this paper for anyone to use. (Round 2 also clears browser state between runs, thanks to Codex.)

Send us a task

The ask is small. Send us one task your customers try to complete with an assistant. It goes in the next matrix, and in three months you'll know who got the booking.

tom@searchleaders.ai

Three research appendices sit behind this paper: a review of LifeVal and twelve agent benchmarks with the gap table; a profile of each assistant's browser and its stated policy on credentials and payment, as documented on 3 November 2026; and a note on agent commerce protocols and the bot management of the Australian airline and travel sites. Every outside figure quoted here carries its source and date in those appendices, and anything we couldn't confirm against a primary source is marked there as unverified and isn't relied on here.

ABOUT

AI Search Leaders is the community for senior people working out AI search, founded by Jeremy Cabral and Tom Critchlow. Jeremy, co-founder and former COO of Finder, led this research and is the paper's lead author. Tom is the founder of the SEO MBA.

CITE THIS PAPER

Cabral, J. and Critchlow, T. (2026). *Task gain: the future of AI search is task completion.* Round 1, 4 November 2026. AI Search Leaders. searchleaders.ai/research/task-gain