AVIAN

The illusion of certainty: AI and government information

We asked three AI tools 102 questions about federal government services and checked all 306 answers against the official source. Only one in five was reliable enough to act on.

By Lauren Leamon

12 November 2026



Contents

Executive summary 3

Method and measures 5

Glossary 7

1 AI can't be relied on when it comes to government information 8

Only one in five answers was reliable enough to act on 8
No tool was consistently reliable 10
The most visible answer isn't necessarily the most reliable 10

2 The problem isn't just wrong information 13

The deciding detail was often missing 14
Some mistakes had serious consequences 14
Caveats didn't help people identify when answers needed checking 15
Vague language made clear rules sound uncertain 16
Some answers contradicted themselves 18

3 Where AI gets its answers matters 19

Answers relying on official sources were generally more reliable 19
More sources didn't mean better answers 21
When AI looked beyond government websites, source quality varied widely 22
Old information was often presented as current 23
Tools sometimes assumed where a person lived 26

4 Where the risks are greatest 29

Reliability varied substantially by topic 29
Some scenarios defeated all three tools 31
Asking a better question didn't reliably produce a better answer 33
Questions asking for an explanation were more often wrong 36

5 Government information is both the problem and the solution 38

Even official sources sometimes disagreed 38
AI tools often bypassed the agency responsible 39
AI reliability is also a content management challenge 40

6 Implications and recommendations 42

For government agencies 42
Further approaches worth testing 44

How to read these results 46

Notes 47

Appendix: data tables 48

Executive summary

1 in 5 AI answers about government services were reliable enough to act on

AI tools are fast becoming one of the ways Australians look for information, including about government services. People ask chatbots like ChatGPT and Microsoft Copilot to explain payments, tax rules and visa conditions, often expecting a quick, clear answer instead of working through several websites. Increasingly, AI answers also reach people who never chose to use AI at all. Google's AI Overviews now appear at the top of search results by default, so a person looking up a government payment or rule may get an AI-generated answer before they reach an official page.

We wanted to know whether those answers can be trusted. We asked ChatGPT, Google AI Overviews and Microsoft Copilot 102 questions about federal government information across 17 topics, from tax and Centrelink to biosecurity and citizenship, and scored all 306 answers against what the official source says.

Only one in five answers (20.9%) was reliable enough to act on. Almost half (49.3%) were partly reliable, and 29.7% were unreliable. For 51.0% of questions, none of the three tools gave a reliable answer. Checking one tool against another won't fix this.

Results differed considerably between tools. ChatGPT gave the most reliable answers (33.3%), followed by Google AI Overviews (16.7%) and Copilot (12.7%). The tool people are most likely to see without choosing it, Google AI Overviews, was among the least reliable and cited social media in 57.8% of its answers.

The biggest problem wasn't wrong information. It was missing information. Just over half of answers were accurate, but most of those left out key facts, and the missing detail was often the condition, exception or recent change that decided what the person should do. 4.9% of answers contained a high-risk error that could lead someone to miss out on support, breach a legal requirement or make a costly decision.

The signals people rely on to judge an answer weren't dependable. Caveats appeared on strong and weak answers alike, and 44% of unreliable answers gave no hint the person should check. Long lists of sources didn't mean better answers. Asking a more precise question helped ChatGPT but made little difference to the other two tools.

Where answers came from mattered. Answers drawing mainly on official sources were far more likely to be reliable (29.0%) than those citing no official source (2.3%). Outdated figures appeared in 14.1% of answers, none of them reliable, and were often traced to commercial websites still carrying old rates, fees and thresholds.

While not as common, government content did also present some issues. In several cases we found official pages that contradicted each other, or that still described programs and rules that had since ended. AI tools repeated these with confidence.

Agencies can't control how AI tools generate answers, but they can shape the information those tools draw on. We recommend agencies test how AI tools answer questions about their services, check that related official pages agree, make changes and closures explicit, state clearly when figures apply, and prioritise content where errors could cause harm. These steps make government information more reliable for everyone, whether they use AI or not.

THE SHORT VERSION

Government agencies need to:

1. Understand how AI tools represent their information.
2. Maintain accurate, consistent and accessible official sources that work for people as well as AI.
3. Test whether changes to their content actually improve the answers AI produces.

See all our recommendations → · page 42

BEFORE THE FINDINGS

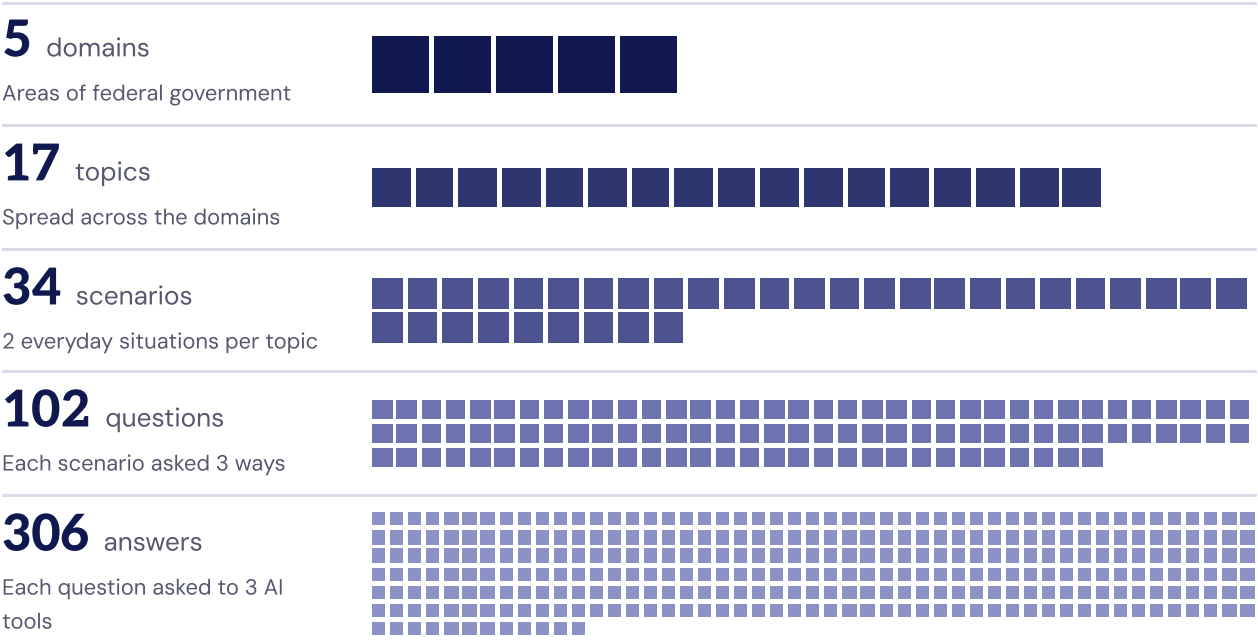
Method and measures

We tested how well three mainstream AI tools answer the kinds of questions people ask about Australian Government information, services and obligations.

FIGURE 1

How 5 domains became 306 answers

From topics to questions to answers



Each square is one item at that level.

HOW WE RAN THE STUDY

1 Choosing the scenarios

The study covered federal government information across 5 domains and 17 topics: financial, health and medical, work and employment, citizenship and identity, and primary industries and energy. Within each topic we wrote 2 scenarios, giving 34 in total. Each scenario was a common thing people do with federal services and had a clear, bounded answer on an official federal website.

2 Designing the questions

Each scenario was asked three ways, reflecting how much the person already knows. An **Unaware** question describes the problem without naming the program. An **Aware** question names the program but not the details. A **Hyper aware** question uses official terminology. Because all three questions cover the same scenario, differences between them show the effect of how people ask. Each question had its own answer card, since a more specific question calls for more specific detail. This gave 102 questions, which we also classified by type (understanding, eligibility, deciding, requirements, process and finding a resource).

3 Testing the AI tools

We tested ChatGPT, Google AI Overviews and Microsoft Copilot, focusing on the free consumer experience. Every question was run once in each tool, giving 306 answers. Responses were collected on 9 October 2026 (and 20 October for the battery rebate questions), using Otterly.AI, an AI search monitoring tool, with location set to Australia. Each question was entered exactly as written, and we recorded the full response and every cited source.

4 Setting the correct answers

Before testing, we wrote an answer card for each question, based on the official federal source. Each card summarised what the source said and listed 4 to 6 key facts a complete answer should include. We judged accuracy against the official source.

5 Scoring the answers

Each answer was scored against its answer card for accuracy (accurate, mostly accurate, inaccurate), completeness (the share of key facts included), whether it acknowledged uncertainty, how much it relied on official sources, the type of any errors, and their severity (low, medium or high). A key fact stated incorrectly didn't count towards completeness, and was recorded as a wrong or outdated detail rather than as missing information.

6 Rating reliability

We combined these into an overall reliability rating. A **Reliable** answer was accurate, covered at least 80% of the key facts and had no error above low severity. An **Unreliable** answer was inaccurate, contained a high-severity error, or covered less than half the key facts. Everything in between was **Partly reliable**.

7 Finding patterns

To understand why answers succeeded or failed, we looked for patterns in sources, error types and question characteristics. These are associations that point to likely causes, not proof of them.

Glossary

HOW WE RATED ANSWERS

Reliable	Accurate, covering at least 80% of the key facts, with at most a minor error.
Partly reliable	Broadly right, but missing important information or containing an error that could affect what someone does.
Unreliable	Wrong, covering less than half of what the person needed, or containing an error that could cause real harm.

HOW WE ASKED

Unaware	Describes the problem, not the program.
Aware	Names the program but not the details, or says what they want but not how to get it.
Hyper aware	Asks for the specific thing they need, using official terms.

HOW WE RATED SOURCES

Primarily official	All or most sources from government.
Mixed	Government and non-government sources in similar numbers.
Minor	One or two government sources among mostly non-government ones.
No official source	No government source was cited.

HOW WE GROUPED QUESTIONS

Eligibility	Whether the person qualifies, such as whether they can get a payment or rebate.
Understanding	What something is or how it works, such as how a rebate is calculated or how a debt is repaid.
Requirements	What the person needs to have or provide, such as documents or records.
Process	The steps to do something, such as how to apply or register.
Deciding	The rules, costs or consequences to consider before acting.
Find a resource	A specific form, page, tool or contact.

CHAPTER 1

AI can't be relied on when it comes to government information

The promise of AI-generated answers is simple: ask a question, get the information you need, and move on. No searching through government websites, comparing pages or trying to work out which rules apply to you.

But that convenience depends on something important. The answer needs to be right, and it needs to tell you enough to make a decision.

Our testing found a substantial gap between those two things. AI tools frequently produced answers that sounded helpful and contained correct information, but left out details that could change what someone should do. Sometimes the answers were simply wrong. And the usual signs of trustworthiness — confident language, caveats and long lists of sources — weren't always useful guides to which answers could be relied on.

Only one in five answers was reliable enough to act on

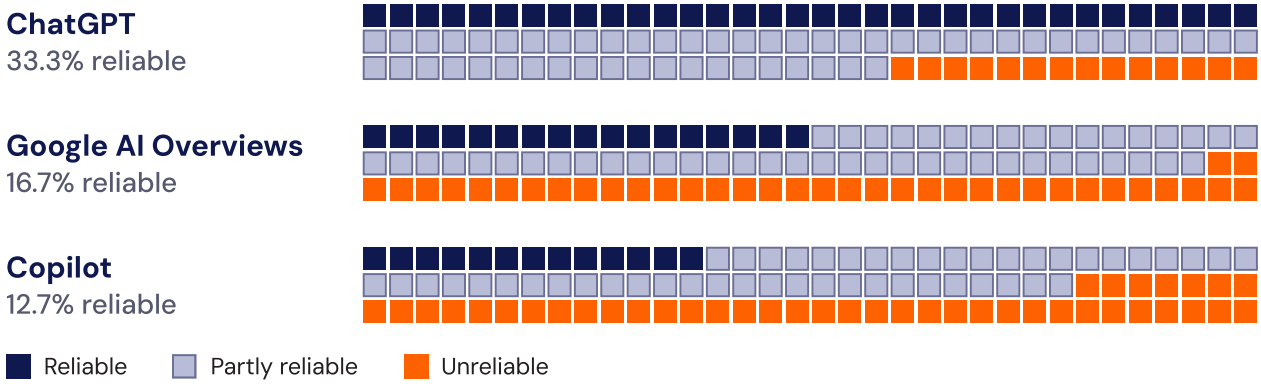
HOW WE RATED ANSWERS

We assessed each answer against three levels of reliability:

Reliable	accurate, covering at least 80% of the key facts, with at most a minor error.
Partly reliable	broadly right, but missing important information or containing an error that could affect what someone does.
Unreliable	wrong, covering less than half of what the person needed, or containing an error that could cause real harm.

FIGURE 2

Each square is one answer. We asked three AI tools the same 102 questions, which gave us 306 answers to check.



Across 306 answers from ChatGPT, Google AI Overviews and Copilot, ■ just 20.9% of answers were reliable enough for someone to act on. ■ Almost half (49.3%) were partly reliable, while ■ 29.7% were unreliable.

The results differed considerably between tools. ChatGPT performed considerably better than Google AI Overviews and Copilot. But even the best-performing tool gave a reliable answer only one-third of the time.

But perhaps the most concerning finding is that the tools often failed on the same questions.

Source: Avian analysis, 306 answers to 102 questions, collected 9 and 20 October 2026.

Data table, page 48

For 51.0% of the questions we tested (52 of 102), not one of the three tools produced a reliable answer. Only a single question, about the ATO's fixed rate method for working from home expenses, received a reliable answer from all three.

This challenges the idea that checking an answer against another AI tool is a useful safeguard. Getting a second opinion doesn't help much if neither answer gives you the information you need.

No tool was consistently reliable

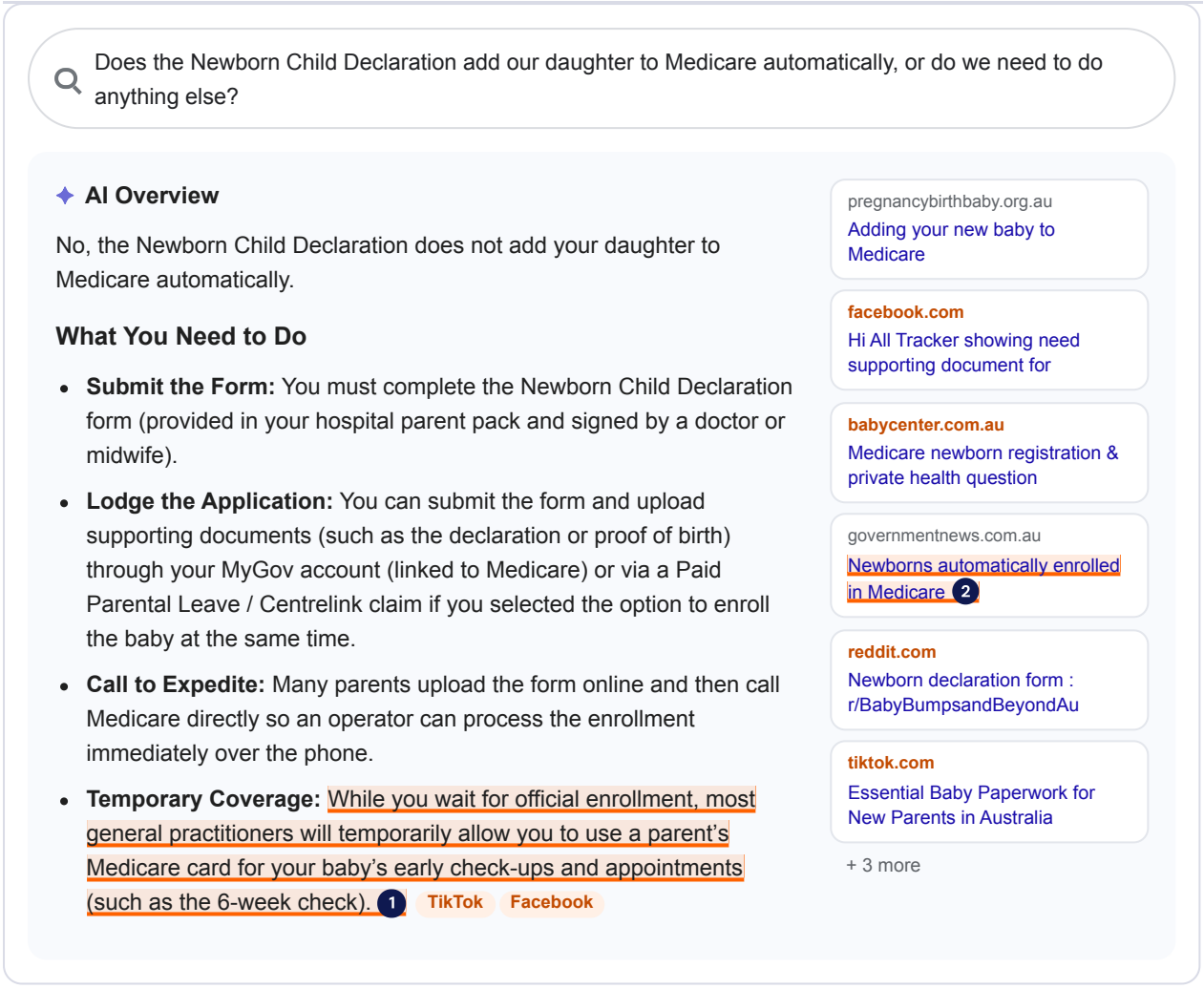
The difference between accuracy and reliability is particularly revealing. ChatGPT's answers were rated accurate 82.4% of the time, yet only 33.3% were complete and dependable enough to act on. An answer could get the facts right and still leave someone without the information they needed.

The other two tools performed substantially worse. Just 16.7% of Google AI Overviews and 12.7% of Copilot answers were reliable. More than a third of answers from each were unreliable, and Copilot gave inaccurate answers almost a third of the time.

The most visible answer isn't necessarily the most reliable

Google AI Overviews presents a particular concern because people don't have to choose to use it.

Google AI Overviews appear automatically at the top of Google search results, unlike ChatGPT and Copilot, which people generally need to choose to use. Google handled 87.9% of Australian searches in September 2026, according to StatCounter,[3] giving AI-generated answers enormous potential reach.



1 **A wrong detail we rated high-risk.** Its only sources for this claim were TikTok and Facebook.

2 **It cites an article that contradicts its own answer.** The headline says newborns are enrolled automatically.

– **No Services Australia or myGov source.** Of nine sources, one was a government site; the rest were forums, social media and a news site.

Recreated from the Google AI Overview answer we collected on 9 October 2026. The answer text is unchanged; inline source links are shortened.

Yet Google AI Overviews produced reliable answers just 16.7% of the time, while more than one in three (35.3%) were unreliable.

Its choice of sources was also striking. Only 5.9% of its answers relied primarily on official government sources. Instead, it frequently drew on social media: 57.8% of its answers cited Facebook, Reddit, YouTube, TikTok or Instagram.

These findings raise a practical concern. People may encounter an AI-generated answer while looking for government information without necessarily recognising that it has been generated differently from a conventional search result. If the answer is incomplete or wrong, they may not know they need to check it.

The risk isn't just that people might trust an incorrect answer. It's that they may have no reason to suspect the answer is incomplete in the first place.

CHAPTER 2

The problem isn't just wrong information

An AI answer doesn't have to be factually wrong to lead someone astray.

In our testing, the most common problem wasn't invented information or obvious mistakes. It was answers that got the general rule right but missed a condition, exception or recent change that determined what someone should actually do.

Other answers gave conflicting advice, made clear rules sound uncertain, or confidently presented information that was out of date.

These failures matter because people don't just use government information to learn something. They use it to make decisions: whether to apply for a payment, register for tax, bring medicine into Australia or challenge a dismissal.

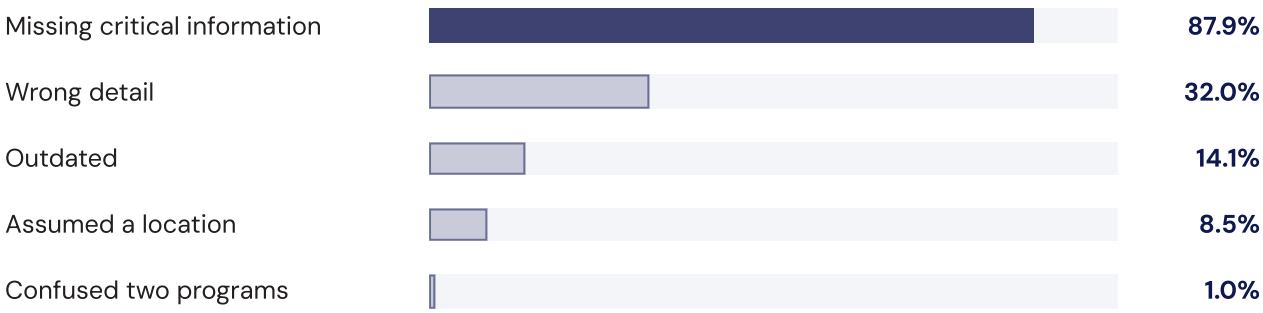
An answer can sound helpful, contain correct facts and still lead someone to make the wrong decision.

FIGURE 3

Missing information was by far the most common problem

% of all 306 answers with each problem

Some answers had more than one problem.



Source: Avian analysis.

Data table, page 48

The deciding detail was often missing

Just over half of all answers (52.9%) were rated accurate, but only 20.9% were reliable. Much of this gap came from information that was missing.

More than half (54.3%) of the answers rated accurate still covered less than 80% of the key facts.

8.8% of answers covered every key fact.

The answers often included extra explanation or information the tool considered relevant. But this does not necessarily help people understand what applies to them or what they need to do next. The missing information was frequently the detail that determined the right course of action.

For example, across the three questions about bringing prescription medicine into Australia, seven of the nine answers left out that Australian residents need an Australian prescription.

That single omission could change whether someone is complying with the rules, even if everything else in the answer is correct.

We found this pattern repeatedly. Answers explained the broad rule but left out the condition that determined whether it applied, the exception that changed the outcome, or an important change to the rules.

The difference between a correct answer and a useful answer was often the one detail that mattered most.

Some mistakes had serious consequences

Not every error has the same consequences. An incorrect fee, a missed eligibility condition or a misunderstanding of a legal requirement can affect someone's finances, entitlements or obligations.

1 in 20 answers contained a high-risk error.

15 OF 306 ANSWERS (4.9%)

GST registration shows how easily this can happen.

CASE STUDY

A new freelancer and GST

Most freelancers must register for GST once their turnover reaches $75,000. But the ATO's registration page also lists earning income through digital platforms, such as websites or apps, as a reason to register, separately from that threshold. Our questions were for a new freelance graphic designer and didn't say how they found clients, so a good answer would have flagged this. None of the nine answers did.

CASE STUDY

Citizenship for a baby born overseas

In another scenario, we asked about citizenship for a baby born overseas to an Australian parent. Where the parent is a citizen by descent, they must have lived in Australia for two years before applying. Two Copilot answers incorrectly said those two years had to be before the child's birth.

A parent who could meet the requirement after their child was born might wrongly conclude their child wasn't eligible.

These aren't simply technical inaccuracies. They are mistakes that can change what someone believes they're entitled to do, or required to do.

Caveats didn't help people identify when answers needed checking

AI tools often acknowledge uncertainty or recommend checking with an official source. These warnings might seem like a useful way to judge how much confidence to place in an answer.

But they weren't a dependable guide to reliability.

ChatGPT acknowledged uncertainty or suggested checking with an agency in 99.0% of its answers, compared with 51.0% for Google AI Overviews and 58.8% for Copilot.

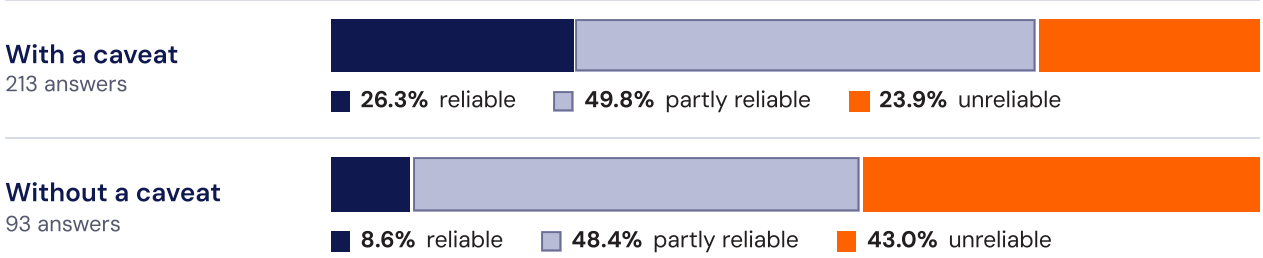
But caveats appeared in both strong and weak answers. While 87.5% of reliable answers included one, so did 56.0% of unreliable answers.

More concerningly, 44.0% of unreliable answers gave no indication that the information should be double-checked. More than half of these came from Google AI Overviews.

FIGURE 4

A warning didn't mean an answer was reliable

% of answers that were reliable, with and without a caveat



Source: Avian analysis, 306 answers.

Data table, page 48

> 44% of unreliable answers gave no warning that they needed checking.

Vague language made clear rules sound uncertain

The problem wasn't only whether tools included caveats, but how they used them.

ChatGPT, the tool that caveated most, often used words like "generally", "may" and "some people" in a way that turned definite rules into apparent judgement calls. For example:

> When the question said the person was getting a Centrelink payment and had started café shifts, ChatGPT said they **"generally need to tell Centrelink"**. Reporting income is required, and failing to do so can create a debt.

> When an Age Pension recipient asked about government help with energy bills, ChatGPT said **"some people"** receiving Centrelink payments **"may receive"** the Energy Supplement, and that eligibility depended on their circumstances. The supplement is paid as part of the Age Pension.

ChatGPT

I'm getting a Centrelink payment and I've just started a few shifts at a cafe. Do I need to tell Centrelink?

Yes. If you've started doing paid shifts at a cafe while receiving a Centrelink payment, you generally need to tell Centrelink [1] about the work and report the income you receive. Your payment may change depending on how much you earn and which payment you're getting. Services Australia

What you need to do depends on your Centrelink payment:

- **If you're required to report income** (for example, JobSeeker, Youth Allowance, Austudy, Parenting Payment, some other income support payments): report your gross income (before tax) from the cafe during your reporting period. Services Australia
- **If you've just started work:** you may need to update [1] your employment details so Centrelink knows you have an employer and can set up the right reporting arrangements. Services Australia
- **Don't wait until tax time:** Centrelink uses your reported income to work out your correct payment amount. Services Australia
- You can usually update it through your myGov Centrelink account or the Express Plus Centrelink app. Services Australia

If you tell me which Centrelink payment you're on (JobSeeker, Youth Allowance, Austudy, Parenting Payment, DSP, etc.) and whether the cafe is casual/part-time, I can explain exactly what you'll likely need to report and when.

Ask anything

1 **Hedged wording for a firm rule.** Reporting income isn't optional for people on these payments.

– **What it left out:** that not reporting can create a debt, when to report (every 2 weeks, or within 14 days), and that working doesn't necessarily stop the payment.

Recreated from the ChatGPT answer we collected on 9 October 2026. The answer text is unchanged; source links are shown as labels.

Even answers that covered every key fact could leave people uncertain.

ChatGPT's answer about how an Australian citizen could bring her husband to Australia covered every key fact we assessed. But it used the word "generally" eight times, including to say that she could "generally sponsor" him.

> Each qualification might seem harmless on its own. Together, they made a relatively straightforward process sound much less certain than it was.

This was particularly noticeable in eligibility questions. They had the highest rate of caveats (84.8%) and were among the least complete, covering 57.2% of key facts on average.

Rather than explaining the rules, tools often said eligibility depended on the person's circumstances without setting out what those rules were.

A caveat is useful when someone's circumstances genuinely affect the answer. But when a tool adds one to almost every response, it stops being a useful signal.

The most useful answers were confident about what was certain and cautious only where the answer genuinely depended on the person's situation.

Some answers contradicted themselves

Sometimes the right information was present, but the answer still managed to be misleading.

> One question asked whether someone who was behind on their mortgage could access their superannuation. ChatGPT initially said that falling behind on mortgage payments **"generally does not qualify"** for early release. Only later did it explain that super can be released on compassionate grounds to prevent a lender from selling your home.

> Another answer about aged care costs said people contribute towards personal care, then later said personal care was fully funded.

These contradictions are particularly problematic when the opening statement is wrong or misleading.

People have always skimmed web pages, which is why good content design puts the most important information first. But AI answers raise the stakes. People often use these tools specifically to avoid having to read and reconcile information themselves.

When the deciding detail appears several paragraphs down, or contradicts the opening statement, the answer has failed at the very task people are relying on it to perform.

On an official government website, an agency controls the structure of its information, including how rules, conditions and exceptions are presented. An AI tool rebuilds that structure, often drawing on multiple sources. A confident opening statement can appear just as authoritative as the government information it was based on.

The person reading it has little way of knowing whether they're seeing an established rule, an incomplete summary or the tool's own interpretation.

This is what makes incomplete answers so difficult to recognise. They can be clear, confident and largely correct, while leaving out the one detail that changes what someone should do. The more convincing the answer, the less reason someone has to question what's missing.

CHAPTER 3

Where AI gets its answers matters

An AI answer can look thoroughly researched. It might cite a dozen websites, link to government agencies and provide a detailed explanation. But none of that guarantees the information is right.

Our testing found that where AI tools got their information mattered considerably. Answers drawing on official government sources were generally more reliable. Those relying on commercial websites, social media or outdated information were more likely to be wrong or incomplete.

But the relationship wasn't straightforward. More sources didn't necessarily produce better answers, and even official government information could be misunderstood or out of date.

Answers relying on official sources were generally more reliable

Answers that drew primarily on official government sources were much more likely to be reliable than those that didn't.

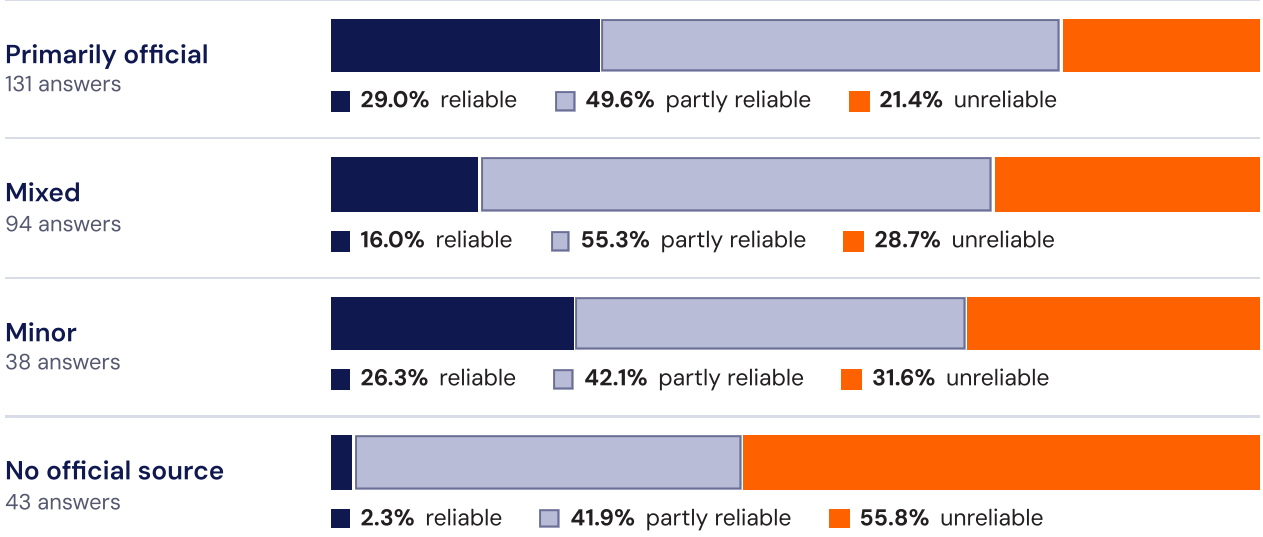
HOW WE RATED SOURCES

We rated each answer by how much it relied on official sources: **Primarily official** (all or most sources from government), **Mixed** (government and non-government sources in similar numbers), **Minor** (one or two government sources among mostly non-government ones) or **No official source**. Government sources included any Australian government website, not only the agency responsible.

FIGURE 5

Answers that cited no official source were rarely reliable

% of answers that were reliable, by how many of their sources were official



Source: Avian analysis, 306 answers. "Minor" means one or two government sources among mostly non-government ones.

Data table, page 49

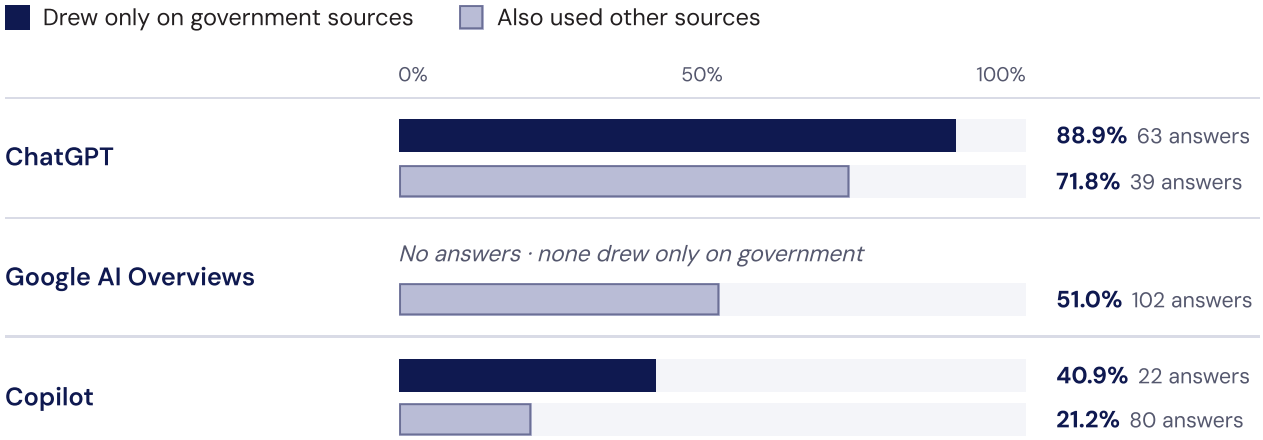
Of the 131 answers that relied mainly on government information, 29.0% were reliable. By comparison, just 2.3% of the 43 answers that cited no official government source were reliable. More than half of those answers were unreliable.

The relationship was also apparent within individual tools. For ChatGPT, 88.9% of answers drawing entirely on government sources were accurate, compared with 71.8% of answers that also used other sources. For Copilot, 40.9% of answers drawing entirely on government sources were accurate, compared with 21.2% of those that also used other sources.

FIGURE 6

Within each tool, answers drawing only on government sources were more accurate

% of each tool's answers that were accurate



Source: Avian analysis. "Only government sources" means the answer cited no non-government source.

WORTH KNOWING

Interestingly, in the instances where Copilot drew only on government sources but still wasn't accurate, the reasons included mislabelling phone lines and apps, misstating limits, turning a requirement into an option, and repeating government pages that were themselves out of date or contradictory.

The absence of official information was a particular warning sign. Fewer than half (46.7%) of answers containing a high-risk error cited the relevant government agency, compared with 94.7% of answers without errors.

But citing government information was no guarantee of getting the answer right.

Copilot sometimes cited government pages alongside outdated figures from commercial websites, or misapplied the rules on the pages it cited. Identity and documents answers drew more heavily on the relevant agency than any other topic, yet only 11.1% were reliable.

The findings suggest that official sourcing matters, but it is not enough on its own to guarantee a reliable answer.

More sources didn't mean better answers

A long list of citations can make an answer look convincing. But the number of sources told us surprisingly little about whether an answer was accurate.

ChatGPT cited the most sources, averaging 14.6 per answer, followed by Google AI Overviews at 9.5 and Copilot at 4.6.

But across tools, the number of sources made little difference to accuracy.

For ChatGPT, answers with ten or more sources were accurate 80.3% of the time, compared with 88.5% for answers with fewer sources. Google AI Overviews showed a similar pattern: 44.9% and 56.6%, respectively.

> The type of source had a larger impact on accuracy. Accurate answers cited an average of 8.0 government sources. Inaccurate answers cited just 2.5.

Accurate answers also cited fewer non-government sources (2.8 on average) than inaccurate answers (4.5).

A long list of sources can make an answer look convincing but the number tells a reader little about whether the information is authoritative, current or relevant to their circumstances.

When AI looked beyond government websites, source quality varied widely

More than a third of all sources (35.3%) were non-government sources, drawn from hundreds of different domains.

These ranged from specialist legal and financial advice to commercial calculators, comparison websites, blogs, social media posts and online forums. The most frequently cited included Facebook, YouTube, Reddit and Instagram.

However, non-government sources aren't a problem in and of themselves. Commercial and specialist websites can contain useful explanations of government programs and policies. They can, of course, also contain outdated figures, incomplete guidance or information that had little to do with the question.

Of course, these websites aren't necessarily responsible for maintaining the authoritative version of a rule, payment rate or eligibility requirement. Their information may be accurate when published but become outdated as policies change.

AI tools can bring all these sources together into a single response, without making the differences in authority or currency apparent to the person reading it.

An answer based on a current government page and one based largely on commercial explainers or social media posts may look equally convincing.

> Google AI Overviews showed this pattern particularly clearly, citing at least one social media platform in 57.8% of its answers.

Old information was often presented as current

Government information changes frequently. Payment rates are indexed, thresholds change each financial year, and programs open, close or change their rules.

In our testing, many errors involved figures that had been correct months or a year earlier.

Keeping information accurate and consistent across government websites and services is an ongoing information governance challenge. Payment rates, thresholds and program details often need to be updated across multiple sources.

We found cases where outdated figures remained published on official government websites alongside current information. Some of these older figures were difficult to find through normal website navigation but could still be retrieved by AI tools.

> This highlights the need to consider not just how government information is maintained, but how easily AI tools can find and distinguish current, authoritative information from outdated content.

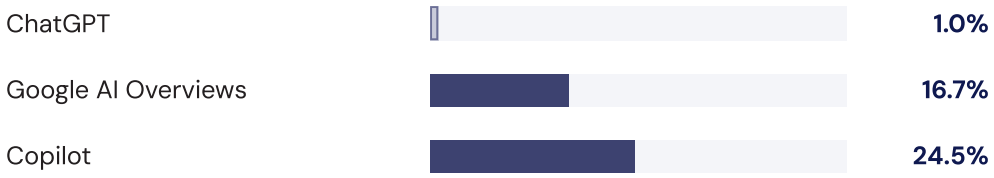
Outdated information appeared in 14.1% of answers, across 20 of the 34 scenarios. None of these answers was rated as reliable.

Answers containing outdated information were also less likely to include a caveat. This was most noticeable in Google AI Overviews: 29.4% of answers with outdated information included a caveat, compared with 55.3% of its answers without outdated information.

FIGURE 7

Outdated information came almost entirely from Copilot and Google AI Overviews

% of each tool's 102 answers that gave an old figure or rule



Source: Avian analysis.

Outdated information came almost entirely from Copilot and Google AI Overviews. Copilot used an outdated figure or rule in 24.5% of its answers, and Google AI Overviews in 16.7%. ChatGPT did so once.

EXAMPLES OF OUTDATED FIGURES

Some figures were only months out of date, including JobSeeker rates and aged care fees from before the September 2026 indexation. Others were years old. Copilot gave the 2023–24 working from home deduction rate and partner visa fee, the 2024–25 unfair dismissal income threshold, and an annual Child Care Subsidy cap that was abolished in 2023.

Some answers described rules that no longer applied. In answer to a question about how many hours an international student can work, Copilot said nurses working in aged and disability care were exempt from the 48-hour fortnightly work limit, even though that exemption ended in 2023. Someone relying on this answer could breach their visa conditions.

Outdated information was particularly common in care costs, study loans and agriculture, where figures are regularly indexed or programs change.

Information that was once correct could be presented as current advice, with no indication that the rules had changed.

Commercial websites helped keep old figures in circulation

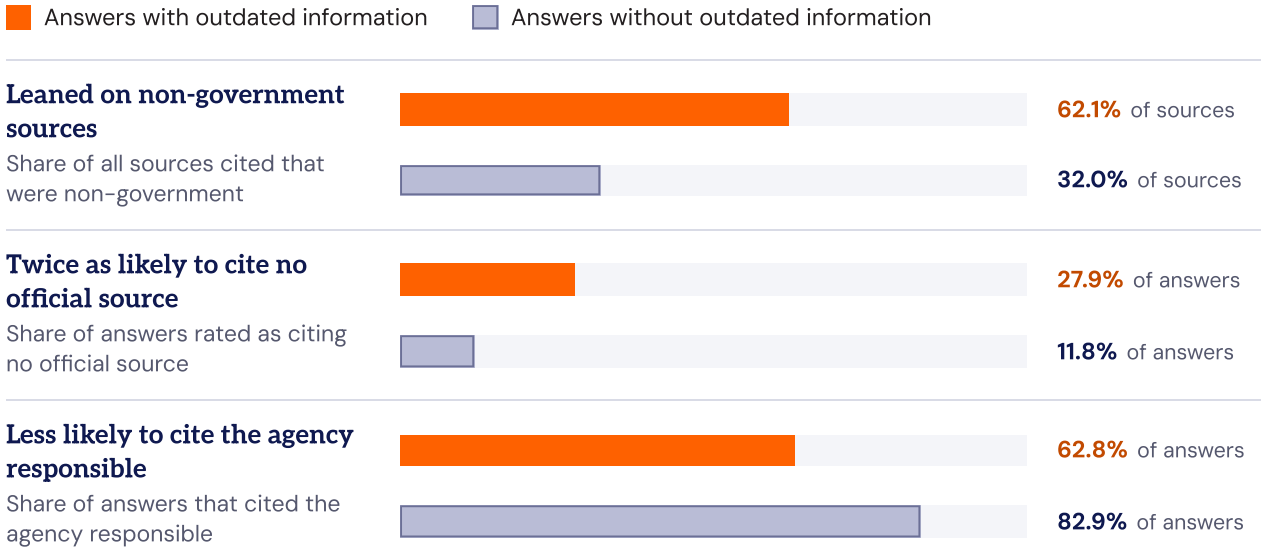
When we examined where outdated information came from, commercial websites appeared repeatedly.

In answers containing outdated information, 62.1% of cited sources were non-government websites, compared with 32.0% for other answers. These answers were also more than twice as likely to cite no official government source at all.

FIGURE 8

How answers with outdated information differed from other answers

Where the answers got their information (43 answers with outdated information, 263 without)



Source: Avian analysis.

Several examples stood out:

Pay rates	Copilot's three answers about a 19-year-old's retail pay all cited fairworkmate.com.au, a commercial pay calculator whose name closely resembles that of the Fair Work Ombudsman. Of the 12 answers across all tools that cited it, eight contained wrong or outdated details.
Student loans	Every answer that cited hecsdebtcalculator.com.au gave outdated repayment figures, including a VET Student Loan threshold of $54,435, the 2024–25 figure, rather than the stated figure of $69,528.
Aged care	Every answer citing agedcaredecisions.com.au or superiorcare.com.au gave outdated fees.
Citizenship	Copilot gave the citizenship by descent fee as $370, which we traced to a migration law firm's page that still showed the old fee rather than the current fee of $380.

Some commercial websites even included "2026" in their page titles while still displaying figures from previous years. The date in a page title was therefore not a reliable indication that the information was current.

And the problem wasn't confined to third-party websites. We also found older government pages that remained available without clearly identifying their information as superseded. In one case, Copilot cited a current Child Care Subsidy guide but still used the previous year's figures.

Outdated information was more common when questions gave more context, such as naming a specific program. Questions like these tended to prompt tools to give specific rates, fees and thresholds, which is where old figures were often surfaced. This suggests that giving the AI more detailed prompts or information does not necessarily lead to more up-to-date answers.

These examples show how difficult it can be to distinguish current information from older guidance once both are available online.

A figure can remain published long after it stops applying. AI tools can retrieve it, combine it with newer information and present the result as if it were current.

And when a tool cannot reliably distinguish the authoritative, current information from everything else available online, an answer can be extensively sourced and still be wrong.

Tools sometimes assumed where a person lived

Sometimes, AI tools went beyond answering the question and filled in details the person hadn't provided.

None of our questions specified a state or territory. Even so, ChatGPT assumed a location in 14.7% of its answers and Copilot in 10.8%. Google AI Overviews did not make these assumptions, instead listing options by state or asking where the person lived.

1 of 26 answers that assumed a location were reliable.

The tools didn't make consistent assumptions. ChatGPT placed people in New South Wales, Victoria, Queensland, Western Australia and South Australia. Copilot assumed Queensland, Victoria, New South Wales and Western Australia. Across the three energy bill questions alone, the tools picked four different states across six answers.

In the most extreme case, ChatGPT answered a question about Australia's federal battery rebate with information about a US tax credit. It said the rebate had expired and cited only American sources. An Australian household could consequently miss out on a discount of around 30%.

ChatGPT

How does the federal battery rebate work if we already have rooftop solar on our house?

If you already have rooftop solar, you generally would have been able to add a home battery and claim a federal tax credit without replacing your solar panels. However, there's an important change for 2026: the U.S. federal residential clean energy tax credit expired for new installations after December 31, 2025. 1 irs.gov uscode.house.gov

The answer depends on which incentive you mean by the "federal battery rebate."

...

What does that mean in 2026?

- **Buying a battery yourself:** The former 30% residential federal tax credit is no longer available for a battery installed in 2026. 1
- **Already installed your battery in 2025:** You may still be eligible to claim the credit on your 2025 tax return, subject to the rules.
- **Other incentives:** State, utility, or local battery rebates may still be available, depending on where you live.

Let me check which incentive applies to you

One quick question: What state do you live in? 2

Sources irs.gov uscode.house.gov starone.org nuwattenergy.com solarhubguide.com

Ask anything

1 **Wrong country.** It answered about a US tax credit and said it had expired. The Australian Cheaper Home Batteries Program was still running.

2 **It asked which US state the person lives in.**

– **Every source was American:** the IRS, the US Code, a US credit union, a New Jersey installer and a US solar blog. No Australian source at all.

Recreated from the ChatGPT answer we collected on 20 October 2026. The answer text is unchanged; some text is removed (marked ...) and source links are shown as labels.

WHEN AND HOW TOOLS ASSUMED A LOCATION

Location assumptions clustered around questions about state-run support, including energy concessions, disability support, agriculture, rent relief and study and training.

The trigger was usually a vague, help-seeking question. Sixteen of the 26 assumptions came from questions written as if the person didn't know which program they needed, and 20 came from eligibility questions. Questions such as "Is there anything to help?" or "Is there any payment I could get?" encouraged tools to search broadly for possible support and then make the answer concrete by selecting a state.

ChatGPT usually signalled its assumptions with phrases such as “If you’re in Victoria” or “Assuming you’re in Sydney”, and often asked for the location to be confirmed. Copilot did not. In all 11 of its location assumptions, it either stated the location as fact or answered for one place without making the assumption clear.

Flagging an assumption is better than presenting it as fact, but it does not solve the underlying problem. The answer may still be built around the wrong state’s schemes, and can blur the distinction between state and federal support. In one answer, ChatGPT placed the Regional Investment Corporation’s national drought loan under the heading “If your farm is in Victoria”, suggesting that it was only available there.

As tools attempt to provide a helpful, specific answer, they can become misleading, filling answers with information the questions did not include. AI researchers describe some questions as “underspecified”, because they leave out details that affect the answer, and have found that AI tools tend to make “incorrect assumptions” to fill those gaps rather than ask.[4]

CHAPTER 4

Where the risks are greatest

Not every question about government information carries the same consequences. Getting an old application fee by a couple of dollars might be inconvenient. Misunderstanding a visa condition, tax obligation or entitlement to financial support can be much more serious.

Our results also showed that some questions were considerably harder for AI tools to answer reliably than others. The topic, the information being requested and even how the question was phrased all made a difference.

Reliability varied substantially by topic

Reliability varied more between topics than between domains, tools, question types or awareness levels.

FIGURE 9

Citizenship and identity fared best, health and medical worst

How reliable the answers were in each domain

Each topic had only 18 answers, so treat topic results as a rough guide.

DOMAIN AND TOPICS	RELIABLE	PARTLY RELIABLE	UNRELIABLE
Citizenship and identity	**29.6%** 16 of 54	**48.1%** 26 of 54	**22.2%** 12 of 54
Visas and immigration	**38.9%** 7 of 18	**38.9%** 7 of 18	**22.2%** 4 of 18
Citizenship	**38.9%** 7 of 18	**38.9%** 7 of 18	**22.2%** 4 of 18
Identity and documents	**11.1%** 2 of 18	**66.7%** 12 of 18	**22.2%** 4 of 18

DOMAIN AND TOPICS		RELIABLE	PARTLY RELIABLE	UNRELIABLE
Financial		22.2% 20 of 90	52.2% 47 of 90	25.6% 23 of 90
Individual tax		38.9% 7 of 18	33.3% 6 of 18	27.8% 5 of 18
Superannuation		22.2% 4 of 18	55.6% 10 of 18	22.2% 4 of 18
Centrelink payments		22.2% 4 of 18	61.1% 11 of 18	16.7% 3 of 18
Care costs and subsidies		16.7% 3 of 18	66.7% 12 of 18	16.7% 3 of 18
Business tax		11.1% 2 of 18	44.4% 8 of 18	44.4% 8 of 18
Work and employment		20.4% 11 of 54	48.1% 26 of 54	31.5% 17 of 54
Study and training loans		33.3% 6 of 18	44.4% 8 of 18	22.2% 4 of 18
Support for new parents		22.2% 4 of 18	50.0% 9 of 18	27.8% 5 of 18
Workplace rights		5.6% 1 of 18	50.0% 9 of 18	44.4% 8 of 18
Primary industries and energy		18.5% 10 of 54	48.1% 26 of 54	33.3% 18 of 54
Biosecurity		22.2% 4 of 18	61.1% 11 of 18	16.7% 3 of 18
Agriculture		16.7% 3 of 18	44.4% 8 of 18	38.9% 7 of 18
Energy		16.7% 3 of 18	38.9% 7 of 18	44.4% 8 of 18
Health and medical		13.0% 7 of 54	48.1% 26 of 54	38.9% 21 of 54
Medicine and health products		16.7% 3 of 18	55.6% 10 of 18	27.8% 5 of 18
Medicare and costs		11.1% 2 of 18	55.6% 10 of 18	33.3% 6 of 18
Disability support		11.1% 2 of 18	33.3% 6 of 18	55.6% 10 of 18

Reliable Partly reliable Unreliable

All questions: 20.9% reliable

Source: Avian analysis, 306 answers.

Health and medical questions performed worst at the broad domain level, with just 13.0% of answers rated reliable and 38.9% unreliable. They also had the highest rate of caveats, at 85.2%.

Some topics fared much worse than others. Only 1 of 18 workplace rights answers (5.6%) was reliable, and business tax, disability support, identity and documents, and Medicare and costs were not far behind, at 2 of 18 (11.1%) each. In business tax, the problems centred on GST: no tool gave a reliable answer about registering, a topic where the ATO's own guidance is unclear. Disability support answers were also the most likely to be unreliable (10 of 18, 55.6%), and had the equal highest share of high-risk errors.

Citizenship, individual tax, and visas and immigration had the highest reliability rates among the individual topics, at 7 of 18 answers (38.9%) each. Even in these areas, less than half of answers were reliable.

> **HOW TO READ THIS**
>
> These results reflect the questions and scenarios tested in this study. They should not be read as a definitive ranking of every government topic or every possible AI interaction. They do, however, show that people asking different kinds of questions of government may face quite different levels of risk.

Still, the variation matters. A government agency cannot assume AI tools represent its information reliably just because the tools perform reasonably well on other government topics.

Some scenarios defeated all three tools

Seven of our 34 scenarios failed to produce a single reliable answer, despite each being asked three different ways across all three tools.

These weren't obscure questions. They covered everyday situations, including someone losing their job, bringing prescription medicine into Australia, checking their pay, seeking disability support and registering for GST.

THE SEVEN SCENARIOS

JobSeeker after losing a job 0 of 9 answers reliable	**What the answers missed:** The liquid assets waiting period and the hardship waiver. Two answers built around a state the person hadn't mentioned.
Specialist gap fees and rebates 0 of 9 answers reliable	**What the answers missed:** That seeing a specialist without a referral means a lower rebate, not none. Government sources disagree on this point.
Bringing prescription medicine into Australia 0 of 9 answers reliable	**What the answers missed:** That Australian residents need an Australian prescription, missing from 7 of the 9 answers.
NDIS eligibility for a child 0 of 9 answers reliable	**What the answers missed:** How the early childhood approach works now. Answers pointed to the wrong programs or to private and state services.
Minimum pay rates for a 19-year-old 0 of 9 answers reliable	**What the answers missed:** Penalty rates. Copilot overstated them in all three answers, while ChatGPT and Google AI Overviews mostly gave the base rate but no penalty rates.
Federal energy bill relief 0 of 9 answers reliable	**What the answers missed:** That the Energy Supplement is paid automatically with the Age Pension. Six of nine answers guessed the person's state.
Registering for GST as a sole trader 0 of 9 answers reliable	**What the answers missed:** When registration is required. One Copilot answer never mentioned the $75,000 threshold, and another wrongly excluded overseas work from turnover. No answer flagged that the ATO lists platform income as a separate reason to register.

In most of these scenarios, the answers covered the general rule reasonably well. They failed on the condition, exception or recent change that determined what the person should do.

These details are often less prominent than the headline information in government sources, or are the details most likely to have changed recently. Their importance can be easy to underestimate, but they may determine whether someone qualifies for a payment, complies with a rule or receives the right service.

Asking a better question didn't reliably produce a better answer

One common piece of advice for getting better results from AI is to write better prompts. Be specific, provide context and use the right terminology.

We tested whether that helped.

HOW WE ASKED

We asked each scenario three ways, reflecting how much the person already knows:

Unaware	describes the problem, not the program. For example, "I work from home a few days a week. Can I claim any of my home costs back?"
Aware	names the program but not the details, or says what they want but not how to get it. For example, "How do working from home tax deductions work for things like electricity?"
Hyper aware	asks for the specific thing they need, using official terms. For example, "How do I use the ATO fixed rate method to claim working from home expenses?"

Because all three questions cover the same scenario, differences between them show how the way people ask affects the answers they get.

Across all three tools, only one thing changed consistently as questions became more precise: answers drew more on the agency responsible. For **Unaware** questions, 39.9% of sources came from the relevant agency. For **Hyper aware** questions, it was 54.7%.

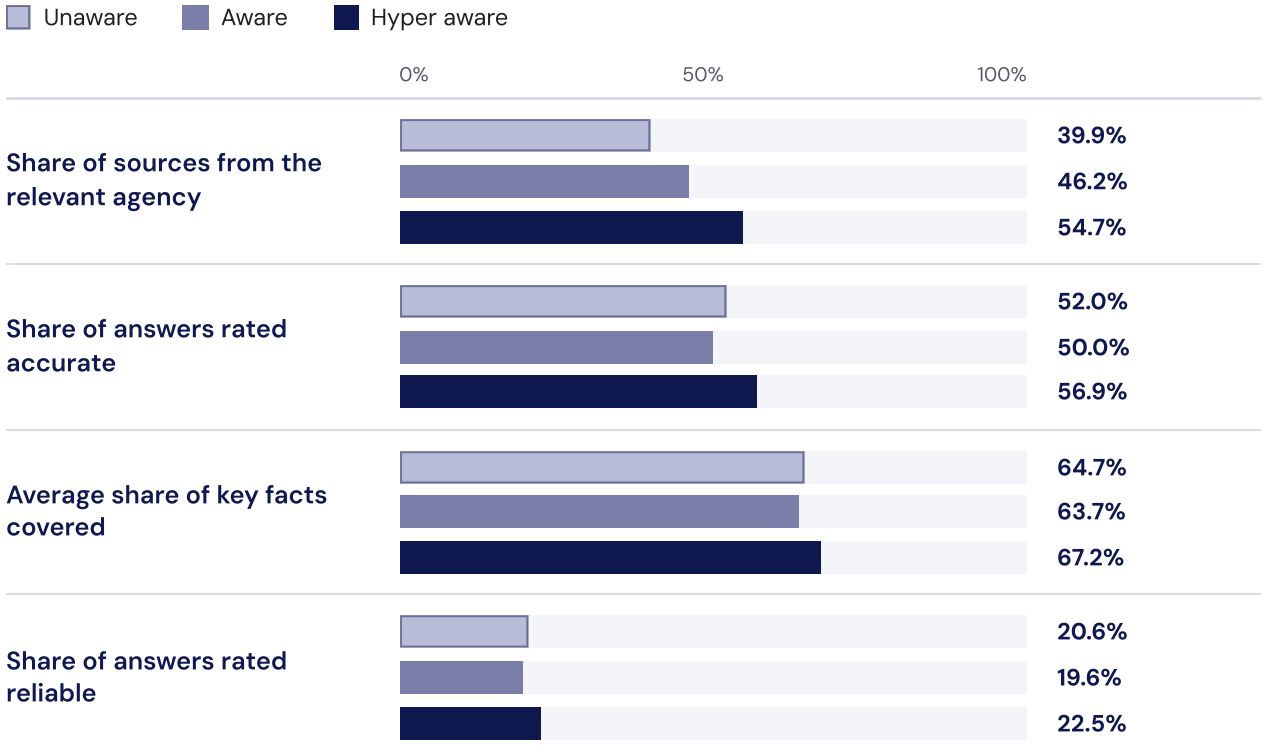
But better sources didn't lead to better answers overall.

Hyper aware answers were only slightly more accurate than **Unaware** answers (up 4.9 percentage points) and slightly more complete (up 2.5 points). The share of reliable answers barely moved, from 20.6% to 22.5%. In individual scenarios, the differences went both ways: **Hyper aware** answers were more complete than **Unaware** answers in 16 scenarios, but less complete in 15.

FIGURE 10

More precise questions found better sources, but barely better answers

All three tools combined, 102 answers for each type of question



Source: Avian analysis. Accuracy for each tool is in the next chart.

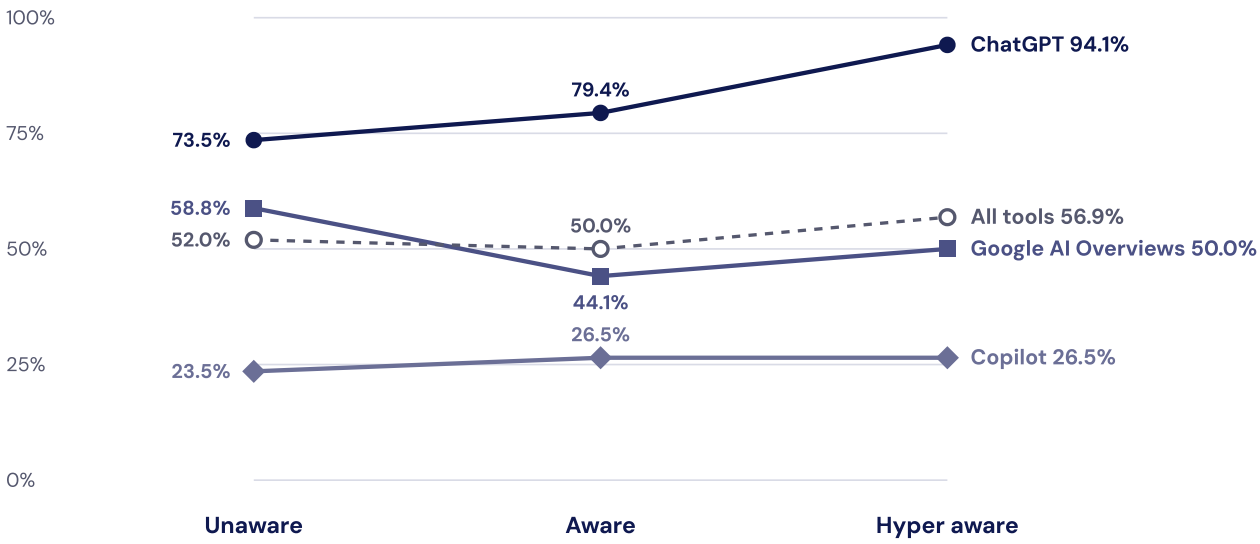
The picture changes when we look at each tool. ChatGPT's accuracy rose sharply, from 73.5% for **Unaware** questions to 94.1% for **Hyper aware** questions. Copilot's barely changed. Google AI Overviews was actually less accurate for **Hyper aware** questions (50.0%) than **Unaware** ones (58.8%).

FIGURE 11

More precise questions only made ChatGPT more accurate

% of each tool's answers that were accurate (34 answers per tool for each type of question)

Each line is one tool. Moving right, the questions get more precise.



Source: Avian analysis.

Data table, page 49

Broad questions presented a different problem. Tools were often cautious about making definite claims, but then failed to identify the relevant program.

WHAT HAPPENED WITH BROAD QUESTIONS

Unaware questions had the fewest inaccurate answers (14.7%), but also the highest rate of caveats (77.5%). When asked for broad help, tools tended to hedge rather than make clearly wrong claims. But they often missed key information or pointed towards the wrong program, latching onto one possible path, such as state workers compensation, state rent relief or private speech therapy clinics, while missing the federal program the person needed.

For example, when the **Unaware** question described a work injury and asked whether any payment was available, answers covered just 16.7% of the key facts on average. When the **Aware** question asked about the Disability Support Pension by name, answers covered 88.7%.

The answers might be strictly correct; but are they helpful?

This creates a problem for people who don't already know how government services work. People often turn to AI precisely because they don't know which program, rule or process applies to them. Yet getting a dependable answer may require knowing what to ask in the first place.

These findings also reinforce the importance of government maintaining clear, accurate and accessible sources of truth. AI can be a useful starting point, but it isn't always a dependable substitute for official information, particularly when someone needs to understand eligibility conditions, exceptions or legal obligations.

People still need somewhere authoritative to go deeper: to explore the rules, understand how they apply to their circumstances and check the details before making a decision.

> **Government information needs to remain useful and navigable in its own right, not just as material for AI tools to summarise.**

Questions asking for an explanation were more often wrong

We also found differences in the kinds of questions people asked.

HOW WE GROUPED QUESTIONS

As well as varying how much the person knew, our questions asked for different kinds of information. We grouped them into six types, based on what the question was asking for:

Eligibility 33 questions	whether the person qualifies, such as whether they can get a payment or rebate.
Understanding 25 questions	what something is or how it works, such as how a rebate is calculated or how a debt is repaid.
Requirements 21 questions	what the person needs to have or provide, such as documents or records.
Process 14 questions	the steps to do something, such as how to apply or register.
Deciding 5 questions	the rules, costs or consequences to consider before acting.
Find a resource 4 questions	a specific form, page, tool or contact.

Deciding and find a resource questions were too few to compare reliably, so the results below focus on the other four types.

Questions asking how something works were almost twice as likely to get an inaccurate answer: 29.3%, compared with 15.6% for other questions.

These were **Understanding** questions, such as how a rebate is calculated or how a debt is repaid (75 answers). The pattern was consistent for all three tools.

Explaining how a program works usually means stating rates, thresholds and rules, and each one is a chance to get something wrong or out of date. These answers also drew more heavily on non-government sources (42.4% of their sources, the highest of any main question type), suggesting that requests for explanations may lead tools towards commercial explainer content.

This content may be easier to understand, but it isn't always current or authoritative.

For example, when asked about the minimum pay for a 19-year-old casual retail worker, Copilot overstated every penalty rate, drawing on a commercial pay calculator. Answers explaining HELP repayments sometimes gave thresholds from the previous financial year, including one sourced to an accountant's newsletter.

The questions themselves weren't necessarily more complicated. But explaining how a government program works requires bringing multiple pieces of information together accurately.

This was a recurring weakness in our study. Tools were often able to find relevant facts, but struggled to combine them into an explanation that was current, complete and correct.

CHAPTER 5

Government information is both the problem and the solution

AI tools can introduce errors even when they find the right government information. They can misunderstand rules, make assumptions or combine facts in ways that change their meaning.

But our testing also uncovered problems in the information available to those tools. Official government pages sometimes contradicted each other, contained outdated guidance or made important details difficult to find.

These aren't problems created by AI. They're longstanding challenges in publishing and maintaining government information. AI makes them more visible, and potentially more consequential, by drawing on information from across different pages and presenting it as a single answer.

Even official sources sometimes disagreed

Official government sources were generally associated with more reliable AI answers. But they weren't always consistent with each other.

In several scenarios, we found government pages that described different rules or gave conflicting information.

> **MBS Online** says a specialist visit without a referral attracts a lower rebate.[1]

> **The myGov specialist page** says the costs won't be covered.[2] Healthdirect, which the myGov page links to, says the same.

Medicare referrals: the three answers in this scenario that we marked inaccurate all said you couldn't get a rebate without a referral, in line with the myGov page.

MORE EXAMPLES OF OFFICIAL PAGES THAT DISAGREED

GST registration	The ATO's main GST registration page lists earning income through the sharing economy or digital platforms under "You must register for GST", without saying whether the $75,000 threshold applies. Its guidance for people providing services through platforms says registration is required once GST turnover reaches $75,000.
Work-related deductions	The ATO's current page says the $1,000 standard deduction applies from 2026–27 returns. An older ATO page still says it's "not yet law".
Biosecurity	The Department of Agriculture's page for international students says home-cooked meals can't be brought into Australia, while its main guidance for travellers says they must be declared for inspection. Its pages also give different maximum penalties for undeclared goods: $3,960 on its airport infringement page and $7,280 on its main traveller page.
Superannuation exit fees	ATO guidance on consolidating super still tells people to check for exit fees, even though these were banned in 2019.

We also found official pages that continued to describe programs or rules that had ended. The federal Energy Bill Relief Fund page still described its 2024–25 extension without clearly stating that the fund had ended in December 2025. A Clean Energy Regulator guide to a carbon farming method did not appear to make clear that the method had closed to new projects.

When official government sources disagree, an AI tool has no reliable way to tell which one is right. It may pick one, blend the two, or hedge between them.

The result can be an answer that matches a government source and is still wrong.

AI tools often bypassed the agency responsible

AI tools were good at finding the big government websites, but often bypassed the specialist agencies that run the program.

This mattered. Answers that cited the agency responsible were reliable 24.5% of the time, compared with 6.6% for answers that didn't, and more than half of those (52.5%) were unreliable.

Major service websites, including Services Australia, Fair Work, the Passport Office, the TGA and the Department of Agriculture, appeared in most or all relevant answers. The gaps were more evident for specialist agencies and topics where federal programs operate alongside state schemes. For example:

Energy bills: Six of the nine energy bill answers cited none of the federal sources covering this support: energy.gov.au, Services Australia's Energy Supplement information or Energy Made Easy. None of the answers to the **Unaware** question did, and Copilot cited no federal government sources at all, relying on state government and commercial sites instead.

VET Student Loans: No tool cited the ATO, even though it manages repayments.

Drought support: Copilot never cited the Regional Investment Corporation, which runs the drought loans.

These gaps didn't necessarily mean the information was unavailable online. But the relevant government source wasn't making its way into the answer.

That matters because the most useful information isn't always on the best-known government website. It may be published by a specialist agency, buried in detailed program guidance or spread across several related pages.

AI reliability is also a content management challenge

Government agencies can't control exactly how AI tools interpret or combine information. But they do have some control over the content those tools encounter.

Our results suggest that this matters. Answers relying more heavily on official sources were generally more reliable, while outdated information and the absence of relevant government sources were associated with poorer outcomes.

For agencies, this creates a challenge that goes beyond writing clear individual webpages.

Information needs to remain accurate as rules change. Related pages need to agree. Closed programs and superseded guidance need to be clearly identified. Conditions and exceptions need to be available alongside the general rule, not scattered across different documents.

Agencies also need to consider whether the information they are responsible for is reaching the tools people increasingly use to find it.

None of this guarantees that AI will generate a reliable answer. We found examples of tools misinterpreting correct information and introducing errors of their own.

But poor content management creates additional opportunities for things to go wrong. If a current payment rate sits alongside an old one, or two official pages describe different eligibility rules, an AI tool has more opportunities to select or combine the wrong information.

The quality of government content has always mattered to the people reading it. Increasingly, it also matters to the AI tools interpreting that content on their behalf.

Making government information current, consistent, complete and easier to find won't eliminate AI errors. But it is one part of the reliability problem that agencies can directly influence.

CHAPTER 6

Implications and recommendations

Only one in five AI answers in our study was reliable enough to act on. Google's AI Overviews, which appear automatically in Google Search, were among the least reliable of the three tools we tested.

Government agencies can't control how these tools generate answers, but they can influence the information available to them. In several cases, we traced unreliable answers to official pages that contradicted each other or hadn't been updated after a change. These are problems agencies can address, and doing so would improve the quality of the information available to people and AI tools alike.

For government agencies

1 Test how AI tools answer questions about your services.

Reliability varied considerably across topics, with between 5.6% and 38.9% of answers rated reliable. Agencies shouldn't assume their information is being represented accurately just because it is up to date and accurate on their website. Testing common questions in the main AI tools, using the kinds of language people actually use, can help identify where answers are inaccurate, incomplete or drawing on other sources.

Once you have results, look for patterns. Are some topics or services answered more reliably than others? Do tools consistently miss the same conditions or exceptions, repeat outdated figures, or draw on the same commercial sites instead of yours? These patterns show where to focus.

Testing is the first step. The next is using those findings to identify where content needs improvement. That might mean making an important exception clearer, updating outdated guidance, resolving contradictions between pages or making authoritative information easier to find.

Some of these approaches overlap with emerging answer engine optimisation (AEO) and generative engine optimisation (GEO) practices, which aim to improve how AI tools find and use online information. But whether particular changes improve AI reliability needs to be tested, rather than assumed.

Finally, test again after making changes. AI tools and government content both change often, so regular testing helps agencies see whether their changes are working and catch new problems early.

2 Keep authoritative information accessible to people, not just AI tools.

AI can help people find and understand government information, but it isn't always a reliable substitute for the original source. People still need somewhere to investigate further, check conditions and exceptions, and understand how rules apply to their circumstances.

Government websites need to remain accurate, accessible and easy to navigate in their own right. That means maintaining clear pathways from general information to detailed guidance, making authoritative sources easy to identify, and ensuring people can find the information they need without relying on an AI-generated summary.

Improving how government content is used by AI matters. But it shouldn't come at the expense of the people who need to read and understand that information for themselves.

3 Check that related official pages agree.

Some unreliable answers reflected contradictions between government pages. Most were within a single agency, such as two ATO pages giving different rules for GST registration through digital platforms, and Department of Agriculture pages giving different penalties for undeclared food at the border. But contradictions also occurred across agencies: MBS Online says a specialist visit without a referral attracts a lower rebate, while the myGov specialist page says costs won't be covered.

When a policy changes, agencies should check all the pages that describe it, including older guidance and information published by other agencies. Updating one page may not be enough if conflicting information remains elsewhere.

4 Make changes, closures and superseded guidance explicit.

When a program ends, a method closes or a rule changes, make that status clear on the relevant pages. Leaving old information available without explaining that it no longer applies creates opportunities for AI tools to repeat it as current advice.

5 Make it clear when figures apply.

AI answers gave different versions of the same rates, fees and thresholds, sometimes drawing on commercial websites with outdated figures. Where amounts change over time, clearly state the period they apply to, such as the relevant financial year. This gives

people and AI tools a better chance of distinguishing current information from older figures.

When figures change, check older pages, factsheets and PDFs that still show previous amounts. Update them, remove them, or clearly label them as historical, so the current figure is the easiest one to find.

6 Prioritise information where errors could cause harm.

Not all content issues carry the same risk. Incorrect information about eligibility for financial support, health costs, employment rights, tax obligations or immigration requirements can have significant consequences for the people relying on it. Agencies should use the potential impact of an incorrect or incomplete answer to help prioritise content reviews and AI testing, particularly where rules change frequently or exceptions determine what someone should do.

Further approaches worth testing

Our findings point to several other ways government agencies might improve how AI tools use their information. These draw on recurring problems in our testing and established good practice for publishing clear, accessible content.

However, we tested AI-generated answers, not the effects of changing government websites. We can't say whether any particular change to content structure, wording or navigation would improve AI reliability.

The following approaches are therefore worth investigating and testing, rather than treating as proven solutions.

Make conditions and exceptions easy to find and understand.

In scenarios where none of the tools produced a reliable answer, the problem often came down to a missing condition, exception or recent change. This included questions about bringing prescription medicine into Australia. Further research could test whether presenting these details more clearly alongside the general rule helps AI tools produce more complete answers.

Clarify where federal and state responsibilities differ.

On topics such as energy bills, drought support and disability services, tools sometimes assumed a person's state or directed them towards the wrong program. Agencies could explore whether clearer distinctions between federal and state schemes, and more explicit links to the responsible agency, help tools identify the right information without making assumptions about a person's circumstances.

Check whether specialist agency information is reaching the right answers.

Our findings showed that AI tools sometimes bypassed the agencies responsible for a particular program. For example, six of the nine energy bill answers cited none of the relevant federal sources, and Copilot did not cite the Regional Investment Corporation in its drought support answers. Agencies could investigate whether their authoritative information is easy for AI tools to find and use. For example, they could check whether:

- key information is on a web page, rather than only in a PDF or behind a login
- page titles and headings use the words people actually search for, not just program names
- the page clearly says which agency is responsible and who the program is for
- major service websites link to the specialist agencies that run related programs, and those agencies link back
- the page is showing up in search results and being cited in AI answers to common questions.

FOR PEOPLE USING AI TOOLS

Treat AI answers about government services as a starting point, not a final answer. All three tools produced unreliable answers in our study, and a confident response didn't necessarily mean the information was correct or complete.

For important decisions, check the details against the responsible government agency's own information, particularly rates, thresholds, eligibility conditions and exceptions. Look for when the information was last updated and whether it applies to your circumstances. Be cautious when an AI answer relies on commercial or social media sources rather than official guidance.

Even official sources can contain outdated or conflicting information. If the rules aren't clear, follow the original guidance, investigate further or contact the responsible agency before making a decision.

AI can make government information easier to access and understand. But when the consequences matter, it's worth going beyond the summary.

ABOUT THE STUDY

How to read these results

LIMITS AND CHOICES

This is an exploratory study, designed to map the landscape and highlight patterns rather than produce statistically representative results. Each scenario has 9 answers, so a single answer moves a scenario's figures by about 11 percentage points. Topic and scenario results are best read as indicative, while the overall and tool-level results, based on 102 to 306 answers, give a firmer picture.

Questions based on common needs	Our team wrote the scenarios to reflect common things people need from federal government services. They show how AI tools handle these everyday situations, rather than representing every question people ask.
First answers to single questions	Each question was asked once, as a single prompt, reflecting the answer someone sees when they first ask. In practice, people may ask follow-up questions, and AI tools can give different answers to the same question on different occasions. Repeating questions to test consistency would be a useful next step.
A snapshot in time	AI tools and government content both change frequently. Results reflect what the tools said on the testing dates, and answer cards reflect what official sources said when we checked them.
Consistent collection	All responses were captured through Otterly.AI, an AI search monitoring tool, so every question was asked in the same way, in a fresh session. What someone sees in their own logged-in session may differ. Where an answer was cut off, such as a truncated AI Overview, we scored what was visible.
Location	Otterly collected responses from within Australia, but we don't know which location each tool detected. So rather than judging whether an assumed location was correct, we recorded where a tool answered as though it knew the person's state when the question didn't say. The variation in assumed states, often between consecutive questions on the same topic, points to inconsistent location signals in the tools.

Three widely used tools	We tested the free versions of ChatGPT, Google AI Overviews and Copilot, the AI tools Australians are most likely to encounter. Other tools and paid versions may perform differently.
Single scorer	One researcher scored all 306 answers against a shared scoring guide and documented edge cases, discussing unclear cases with the team. This kept the approach consistent.
Clear reliability thresholds	The reliability levels use completeness cut-offs (80% and 50%) and severity ratings based on our judgement of what someone needs to act on an answer. Different thresholds would change the proportions in each level and the order of some topics, but not the order of the tools.
Where official sources disagreed	We scored against the answer card, which followed the responsible agency's primary page. Where a rule from that page might not apply to the situation, as with GST registration for platform workers, we treated it as something a complete answer should flag, and scored its absence as a medium-severity omission rather than a high-risk error.
Grouping sources	We grouped non-government sources by type based on their domain names, giving a broad picture of where tools looked for information.
Question type and awareness	Question type sometimes shifts as awareness increases, so the awareness results reflect both what people know and how they naturally phrase their question.
Patterns that point to causes	Source patterns and error types suggest likely reasons for poor answers. Testing whether specific content changes improve AI answers is a natural next step.

Notes

1. MBS Online, Factsheet: Referrals for specialist services, last updated August 2026.
2. myGov, Seeing a specialist, allied health professional or getting screening tests or scans, last updated July 2025.
3. StatCounter, search engine market share in Australia, September 2026. gs.statcounter.com/search-engine-market-share/all/australia
4. Herlihy, C., Neville, J., Schnabel, T. and Swaminathan, A. (2024). On Overcoming Miscalibrated Conversational Priors in LLM-based Chatbots. Conference on Uncertainty in Artificial Intelligence (UAI 2024). https://arxiv.org/abs/2406.01633

APPENDIX

Data tables

FIGURE 2

Answers by tool

	ANSWERS	RELIABLE	PARTLY RELIABLE	UNRELIABLE
ChatGPT	102	34 (33.3%)	54 (52.9%)	14 (13.7%)
Google AI Overviews	102	17 (16.7%)	49 (48.0%)	36 (35.3%)
Copilot	102	13 (12.7%)	48 (47.1%)	41 (40.2%)
All tools	306	64 (20.9%)	151 (49.3%)	91 (29.7%)

FIGURE 3

Missing information was by far the most common problem

Answers with each error type

ERROR TYPE	ANSWERS	SHARE OF ALL ANSWERS
Missing critical information	269	87.9%
Wrong detail	98	32.0%
Outdated	43	14.1%
Assumed a location	26	8.5%
Confused two programs	3	1.0%

FIGURE 4

A warning didn't mean an answer was reliable

Answers with a caveat, by rating

	WITH A CAVEAT	ANSWERS	SHARE
Reliable	56	64	87.5%
Partly reliable	106	151	70.2%
Unreliable	51	91	56.0%

Answers with a caveat, by accuracy

	WITH A CAVEAT	ANSWERS	SHARE
Accurate	134	162	82.7%
Mostly accurate	51	86	59.3%
Inaccurate	28	58	48.3%

Answers with a caveat, by tool

	WITH A CAVEAT	ANSWERS	SHARE
ChatGPT	101	102	99.0%
Google AI Overviews	52	102	51.0%
Copilot	60	102	58.8%

FIGURE 5

Answers that cited no official source were rarely reliable

Answers by reliance on official sources

	ANSWERS	RELIABLE	PARTLY RELIABLE	UNRELIABLE
Primarily official	131	38 (29.0%)	65 (49.6%)	28 (21.4%)
Mixed	94	15 (16.0%)	52 (55.3%)	27 (28.7%)
Minor	38	10 (26.3%)	16 (42.1%)	12 (31.6%)
No official source	43	1 (2.3%)	18 (41.9%)	24 (55.8%)

FIGURE 11

More precise questions only made ChatGPT more accurate

Share of answers rated accurate, by question level

	UNAWARE	AWARE	HYPER AWARE
ChatGPT	73.5%	79.4%	94.1%
Google AI Overviews	58.8%	44.1%	50.0%
Copilot	23.5%	26.5%	26.5%
All tools	52.0%	50.0%	56.9%

AVIAN

Talk to us about testing your content

avian.au/contact