BEARING & COURSE
ADVISORY

POINTS OF VIEW

The model is not the service

A model can outperform a person on a defined measure and still sit inside a less reliable service.

On 14 May 2026, the UK Government officially launched GOV.UK Chat through the GOV.UK app, giving people a way to ask questions in ordinary language and receive answers drawn from official government information.

The service was intended to reduce the need to call a helpline or navigate more than 80,000 pages of guidance. Before opening access more widely, the Government Digital Service had run two public pilots involving more than 10,000 users and 26,000 questions. It tested accuracy, reliability, speed, safety and user trust, and reported that assessed accuracy had improved from an early benchmark of 76 per cent to 90 per cent across the topics tested.

As the service launched, *The Times* reported concerns arising from testing of GOV.UK Chat.

In one reported exchange, the service was asked what would happen to tax-free childcare eligibility if someone's income increased by £1,000 from £99,000. GOV.UK Chat reportedly replied that the person would remain eligible because the scheme did not have an upper income limit affecting them at that level.

The correct individual answer required more information. Eligibility depends on expected adjusted net income rather than salary alone, and the threshold applies when expected adjusted net income is over £100,000. The problem was not simply whether the conclusion was wrong. The reported response omitted the existence of the threshold and did not seek the information needed to determine whether it applied.

The source material was authoritative. The service had been carefully developed and extensively tested. It drew exclusively on published GOV.UK guidance, linked people back to the material it used and had been designed to ask clarifying questions when the original request was ambiguous. The Government Digital Service was also clear that answers falling outside its accuracy measure were often partly correct rather than wholly invented. They answered some aspects of the question but did not meet everything required by the published guidance.

This is a more difficult problem than an obvious hallucination.

An answer can correctly state a general rule while omitting the exception, definition, threshold or personal circumstance that changes what the rule means. It can be grounded in authoritative information, clearly expressed and still be unsafe to act on.

Compare the service, not the answer

Before the wider launch, Ian Murray, the UK minister for digital government and data at the time, told *The Times* that GOV.UK Chat had been "tested to essentially destruction" and was "much more accurate than the human context".

The comparison is revealing because it treats the model's answer and a human answer as equivalent units of work.

They are not produced under equivalent conditions.

A person answering the same question usually operates within a wider service. They may ask for more information, recognise an exception, access details about the person's circumstances, explain uncertainty or refer the matter to someone with greater expertise. Different questions can follow different pathways depending on their complexity and consequence.

The chatbot's accuracy measure assessed the answer it produced. It did not measure all the information, judgement, escalation and accountability surrounding a human response.

The search function that came before it also performed a different role. Search presented sources and left the user to identify the relevant information, compare it with other material and determine how it applied.

A conversational service does more of that work on the user's behalf. It selects information, combines it and presents what appears to be an interpreted conclusion. That is why it can be more useful than a page of search results. It is also why the organisation operating it assumes a different responsibility.

The clearer and more personalised the response becomes, the more reasonable it is for the person to assume that the important interpretation has already occurred. Providing links to the original guidance remains valuable, but it does not return the interaction to search. The service has still selected the information and delivered an answer in its own voice.

GOV.UK Chat warns people that its answers may be inaccurate, simplified or incomplete and requires users to verify information before acting on it. Its algorithmic transparency record describes the service as providing personalised answers based on GOV.UK guidance rather than making official decisions on a person's behalf.

Those safeguards are appropriate, particularly while the service continues to develop. They do not settle what should happen when a reliable answer depends on context the system does not hold.

The service may need to ask another question, explain what it cannot determine, provide only the relevant general information or move the interaction to someone who can understand the person's circumstances.

Government Digital Service research found that some users wanted to speak with advisers who could access their personal information. Work was already under way to provide hand-offs to departmental customer support where required.

That hand-off is part of the capability being designed.

The proper comparison is therefore not between the model and an individual employee. It is between the AI-enabled service and the complete service it is changing.

That comparison includes the speed, consistency and accuracy of the model. It must also include whether the service recognises missing context, routes difficult interactions appropriately and retains accountability for what happens next.

A 90 per cent accuracy result may demonstrate substantial technical progress. It does not show whether the remaining weaknesses occur in low-consequence interactions or in questions affecting someone's tax, payment, eligibility, visa or legal obligations. It cannot determine whether a partly correct answer remains useful or whether the omitted detail changes the conclusion completely.

Accuracy matters, but consequence determines what accuracy, judgement and escalation the service requires.

Human work contains more than answers

The same pattern appears inside organisations.

Most organisations have a documented version of how work happens and a practical version understood by the people doing it.

Experienced staff know which report needs to be corrected before it can be trusted. They know which policy explains the general rule but does not deal with the situation in front of them. They understand that two fields with similar names may represent different things and that a simple request can depend on information held elsewhere.

They also know when the formal process has reached its limit and who can resolve the exception.

These arrangements survive because people compensate for weaknesses in the organisation around them. They compare sources, apply context, identify unusual cases and intervene when an answer does not look right.

Much of that judgement is informal. It may not sit in a document repository, process map, data model or system configuration.

An AI capability can retrieve the approved policy without knowing how experienced staff resolve its ambiguities. It can use the current procedure without knowing that people have developed a workaround to keep the service operating. It can locate the general rule without having the circumstances required to apply it.

Connecting organisational information to a model does not automatically connect the judgement people have been applying around that information.

The system may still produce an impressive answer. It can create a clear and well-structured response that faithfully reflects the organisation's formal records while omitting the practical knowledge required to use them.

This is how AI exposes the organisation underneath it. It makes visible the distance people have been bridging between documented processes and the way work actually happens.

A successful pilot can conceal an incomplete service

Pilots tend to create favourable conditions. The task is constrained, the source material is selected, permissions are limited and specialists are paying attention. Results are reviewed because everyone knows the capability is being tested.

Those conditions help the pilot succeed.

Normal operations contain more variation. People ask unexpected questions. Information changes unevenly. The same term carries different meanings across the organisation. Users have different levels of experience, and the products and models supporting the service continue to change.

A pilot can therefore establish that a model performs a task without demonstrating that the organisation can operate the resulting service reliably.

The useful distinction is between technical performance and operational readiness.

A pilot should test whether the system can produce an acceptable answer. It should also test whether the capability recognises when it lacks the context required to do so and whether the surrounding service responds appropriately when that happens.

Where success depends on curated information, expert review or close supervision, those conditions are part of the capability. Removing them after the pilot does not simply scale what was tested. It creates a different service.

This is where the capability question becomes operational. Assurance, procurement expertise, risk management and operating discipline are not simply mechanisms for approving the technology. They create the conditions under which people can rely on the resulting service.

Readiness belongs to the organisation

Data quality is usually the first weakness raised in discussions about AI readiness. It matters, but GOV.UK Chat shows why it cannot carry the whole question. The information can be official and the response can still lack the context required to apply it.

Global technology consultancy Thoughtworks captures this broader requirement through its FOREST framework: foundational architecture, operating model, ready data, experiences for humans and AI, strategic alignment and trustworthy AI.

The framework's value is in confirming that readiness is an organisational condition. Architecture, data, people, purpose, trust and ownership meet around the work the capability is expected to perform.

Authoritative information does not resolve ambiguity. Strong architecture does not establish ownership. An approved model does not decide where human judgement remains necessary. An assessment does not remain current when the information, integration, supplier or intended use changes.

Continuing ownership matters because AI will not always arrive through an obvious program. A supplier may add summarisation, generated answers or automated recommendations through a routine product update. Staff may gradually use a convenient assistant for more consequential work. What was originally approved can become a different capability without a single clear decision that it should.

Technology may operate the platform, procurement may manage the supplier, legal may advise on obligations and the business area may use the output. That distribution of work is normal. The weakness appears when involvement by several functions is mistaken for end-to-end accountability.

Someone needs to retain a line of sight from the answer the system produces to the organisational outcome that follows. They need to know whether the capability is still fit for purpose, whether its use or consequences have changed and when it should be restricted, redesigned or stopped.

Test the service before scaling it

The concerns reported about GOV.UK Chat do not establish that the service should not have been launched. The public material describes extensive testing, staged implementation, source links, warnings, safety controls and continuing work on customer hand-offs.

The example matters because those measures did not remove the organisational questions. They made them easier to see.

Before scaling an AI capability, leaders should be able to answer three questions.

1. Can the capability recognise when it does not hold the information or context required to produce a reliable answer?

2. Can the service respond appropriately by seeking more information, explaining uncertainty, limiting its answer or moving the interaction to a person?
3. Does someone remain accountable for what happens after the answer is acted upon, including detecting changes to the model, information, supplier or use?

A model may perform better than an individual employee on a defined accuracy measure and still be part of a less reliable service.

Government capability should therefore not be judged only by whether agencies can approve, procure or deploy AI quickly enough.

It should be judged by whether they can design and operate the service that makes its performance trustworthy.