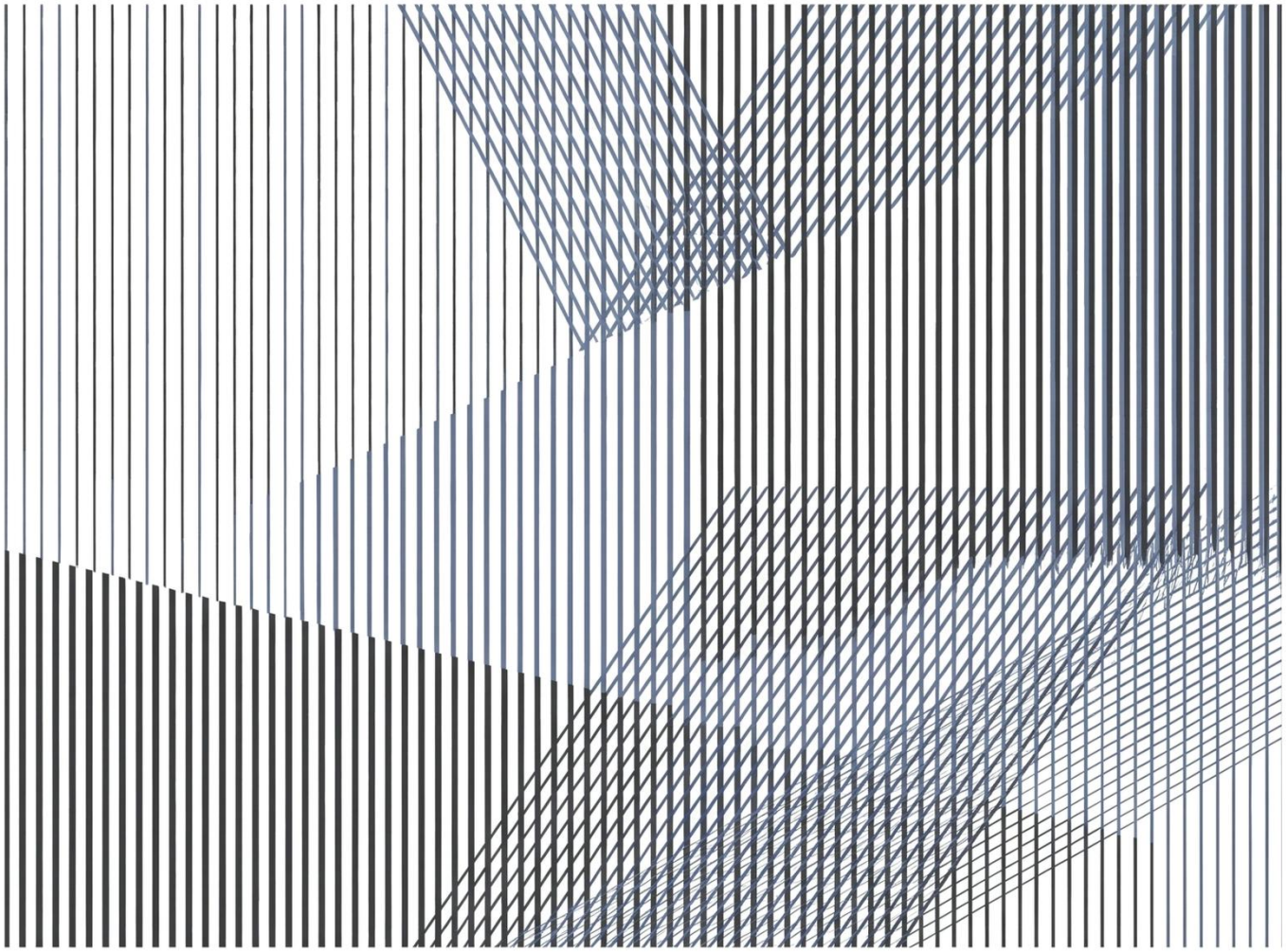
Using AI in Public-Sector Audit

A Guide for Audit Leaders



C.L. Callaghan, M.Sc.
Principal Consultant
November 2026

PolicySpark.ai

ABOUT POLICYSPARK.AI

PolicySpark.ai is an independent consulting practice that helps public-sector organizations use evidence well. Its work covers evaluation, performance measurement, policy research and evidence synthesis for federal departments and agencies and for international development organizations. It combines established research and evaluation methods with AI-assisted tools for finding, organizing and checking evidence, so that every conclusion can be traced to its sources and judgment stays with the people accountable for it.

ABOUT THE AUTHOR

Chris Callaghan is the founder and principal consultant at PolicySpark.ai. Trained in applied science and biology, Chris began in field ecology and ecosystem research. Over more than 25 years in federal service and consulting, Chris has worked in environmental management, public-sector audit, regulatory performance, program evaluation and science policy, bringing scientific, program and stakeholder evidence together into practical advice for decision-makers.

HOW THIS GUIDE WAS PRODUCED

In keeping with the principles set out in this guide, AI tools assisted in preparing it, including organizing a source-linked research database, drafting working text, checking references and consistency, and supporting graphics and layout. The author defined the scope and questions, examined key sources, actively directed and edited every section, and takes sole responsibility for all judgments and conclusions. See the Production note at the end of this guide.

SUGGESTED CITATION

Callaghan, C.L. (2026). *Using AI in Public-Sector Audit: A Guide for Audit Leaders*. PolicySpark.ai.

DISCLAIMER

This guide presents the views of the author. It is intended to inform discussion and decisions within audit institutions and does not constitute professional audit, legal or other advice. PolicySpark.ai uses AI-assisted evidence-synthesis workflows in its own consulting practice; this guide examines source-linked evidence databases as one option alongside simpler approaches. The audit described in Section 3 and Annex A is fictional: the program, figures and outputs were constructed for this guide. Commercial and open-source tools are named as examples of a kind of tool; naming them is not an assessment or an endorsement.

© 2026 PolicySpark.ai. All rights reserved. Public-sector audit institutions may circulate this guide internally for non-commercial use with attribution.

policyspark.ai · chris@policyspark.ai

Contents

Executive summary **5**

1. AI can strengthen public-sector audit, on two conditions **8**

1.1 AI could extend what auditors examine and explain 8
1.2 Convincing output does not establish its own reliability 9
1.3 Audit institutions are experimenting widely, but routine use is proving difficult 11
1.4 Four questions for audit leaders 13

2. What AI can contribute to public-sector audit work **15**

2.1 Finding and preparing evidence 15
2.2 Comparing records, identifying patterns and selecting cases 21
2.3 Developing and testing possible explanations 24
2.4 Communicating findings and analytical work 27
2.5 Coordinating connected tasks 29
2.6 Lessons for leaders across the five kinds of work 31

3. Following one question through a performance audit: an illustrative example **33**

3.1 The audit and its question 33
3.2 Planning: an error in the brief raises the first question 34
3.3 Preparing the records: the headline combines two different things 35
3.4 Fieldwork: providers point to the approval process 36
3.5 Analysis: the explanation changes 37
3.6 Reporting: a check against the verified figures 38
3.7 What the audit institution decides 38
3.8 What the example shows about AI's contribution 39

4. Managing AI use across the institution **41**

4.1 The most promising applications lie where audits currently stop short 41
4.2 Reliance on AI results depends on the task, not only on the tool 43
4.3 Sustaining AI requires continuing investment in people, data and records 47
4.4 Institutions need to know which applications are in use and whether they deliver value 51
4.5 External support is often unavoidable, but responsibility and key expertise stay with the institution 53
4.6 Decisions to keep, extend, change or stop an application require evidence, authority and the ability to act 57
4.7 Disclosure, transparency and holding the institution to its own standard 59

5. Developing the professional capacity to use AI well **61**

5.1 Good assisted work does not show what its author understands 61
5.2 AI could change how auditors learn, for better or worse 64
5.3 Sustaining expertise over a career: leadership choices where evidence is thin 67

6. Synthesis, sequencing and the future of public-sector audit **70**

6.1 The shift beneath the tools 70
6.2 Turning principles into a first-year sequence 71
6.3 Looking ahead: an unpredictable frontier and how public audit could change 72
6.4 Building the profession's own evidence base 74

References **75**

Annex A. The full worked example: using AI across a performance audit **79**

A.1 The program and the audit questions 80
A.2 Building a basis for the inquiry 81
A.3 Following questions through fieldwork 85
A.4 Developing and testing findings 87
A.5 Preparing the report 89
A.6 What the audit institution decides 90

Production note 93

Boxes

Box 1.1. Three audit institutions at different stages of AI use
Box 2.1. Prototypes in Australia and Canada: auditors shape the evidence before AI drafts from it
Box 2.2. An accounting benchmark: errors that accumulated and were then concealed
Box 4.1. IP Australia: capable use of AI with gaps in assurance
Box 4.2. The Australian Taxation Office: governing AI without a central view or advance criteria
Box 4.3. Ontario: unapproved tools and unverified supplier assurances
Box 5.1. Shen and Tamkin (2026): practising with AI and learning less
Box 5.2. Brynjolfsson, Li and Raymond (2025): an AI tool built from the practices of a firm's strongest workers

Exhibits

Exhibit 1.1. How to use this guide by role
Exhibit 2.1. What one record in an evidence database holds
Exhibit 3.1. Each step of the audit changes the next
Exhibit 3.2. Why an aggregate score missed the error in the AI planning brief
Exhibit 4.1. Verification suited to each kind of AI-assisted result
Exhibit 4.2. Proportionate AI operating models by audit office size
Exhibit 6.1. Sequencing leadership action across the first year and beyond
Exhibit A.1. Delivery figures established during planning
Exhibit A.2. AI assistance across the audit
Exhibit A.3. Two illustrative delivery-planning briefs
Exhibit A.4. Assessing the contribution and deciding what to continue

Figures

Figure 1.1. Extending audit reach while supporting conclusions and sustaining judgment
Figure 2.1. How a source-linked evidence database places auditor review before AI drafting
Figure 4.1. Matching verification and control to task stakes up to the auditor judgment boundary

Executive summary

THE MAIN MESSAGE

AI can strengthen public-sector audit by extending what auditors can investigate and explain, on two conditions. Institutions must be able to support the conclusions they draw from AI-assisted work, and auditors must keep the judgment needed to challenge those conclusions.

Why this guide

Public-sector auditors are asked questions that cut across programs and organizations. The evidence needed to answer them sits in documents, data and systems that were never designed to be read together. With limited staff and time, audit teams examine only part of that evidence. A team that has found one credible explanation also tends to settle on it before fully considering what else could account for the facts.

Artificial intelligence (AI) could ease this constraint by helping teams examine more of the relevant evidence and pursue important questions further. However, AI does not change what an audit must deliver: conclusions that the institution can explain and support.

Audit institutions are already exploring AI, for tasks such as anomaly detection, document processing and risk assessment. Moving from pilot projects to routine use is proving difficult, with fragmented data, limited expertise and evolving governance among the obstacles.

This guide is written first for audit leaders who must decide how their institutions use AI, and also serves as a reference for engagement directors, methodology leads and auditors. It concentrates on performance audit, is anchored in Canadian assurance standards, statutory confidentiality and legislative independence, and draws widely on international audit practice and research. It addresses four questions:

1	What can AI contribute to audit work?	→ Sections 2–3
2	How should institutions choose where to use it and decide when its results can be relied on?	→ Sections 4.1–4.2
3	How can worthwhile use be sustained, and when should an application be extended, changed or stopped?	→ Sections 4.3–4.7
4	How can institutions keep developing the professional judgment that AI-assisted work depends on?	→ Section 5

It draws on studies of AI in audit and related professional work, on the published plans and audit reports of audit institutions, and on a fictional performance audit that shows how the pieces fit together. Its focus is using AI in audit work rather than auditing AI systems in government, though governing internal AI use well strengthens credibility when auditing others. Exhibit 1.1 shows which sections fit each reader role so leaders can skim while audit teams use the full reference.

What the guide finds

REACH → Sections 2 and 3

AI's most promising contribution is reach.

Most audit applications reported so far aim at efficiency in existing work. The larger opportunity lies where audits currently stop short. AI can help auditors assemble dispersed evidence while keeping every passage and figure linked to its source. A team can then compare a whole program instead of a few files, select cases with better information and test its conclusions against more explanations where teams normally stop. The potential gain is a better-supported answer, not only a faster one, especially when teams carry what they learn from one step to the next. Wider coverage also carries a risk: in one financial-audit experiment, auditors who tested every transaction did less follow-up when outside evidence later suggested fraud.

RELIANCE → Sections 2.6 and 4.2

How far AI can be relied on depends on the task, not only on the tool.

In experiments, AI assistance has improved speed and quality on some tasks and reduced accuracy on others. Output can be fluent, well cited and wrong, and it does not show what a search left out. The same tool can therefore suit one role and not another: screening leads tolerates more error than shaping audit scope or supporting a finding. Review is not a safeguard by itself, because reviewers can become less careful when they believe a machine prepared the work. After checking results, auditors still decide which errors could change a finding and document consequential checks on the engagement file for entity clearance and legislative scrutiny.

SUSTAINING → Sections 4.3–4.7

Sustaining AI requires continuing investment and independent safeguards.

Once in use, an application depends on people who can maintain it, users who can interpret its results and a record of why it was approved. Recent audits of public agencies show the risks. The Australian Taxation Office lacked a central view of its AI use and baseline comparisons. In Ontario's government, staff accessed about 400 AI websites in five months (60% rated unsafe) and officials left gaps in supplier test reports unaddressed. Public audit offices must also protect confidential and Protected B (or equivalent) records under statutory secrecy, preserve legislative independence on cloud or shared infrastructure and match governance to office size. Whether an application works and whether it is worth its cost are separate questions, and evidence on retiring applications or granting autonomy remains thin.

JUDGMENT → Section 5

Judgment has to be developed deliberately.

A well-reasoned working paper used to signal that its author understood the reasoning. With AI, a supervisor may receive a stronger product without knowing how much its author understands. Delegating work to AI could remove practice and feedback that build judgment, or free time for harder inquiry. Evidence points both ways. In one small experiment, participants who practised an unfamiliar task with AI scored lower on an unaided test. In one firm, an AI tool improved speed and quality for less experienced workers, while the most experienced saw small quality declines. No study reviewed here follows auditors over extended AI-assisted work, so the long-term effect on expertise remains unknown.

What audit leaders can do

The guide recommends six connected actions. The first starts from the opportunity, the next four concern supporting the conclusions and the sixth concerns sustaining judgment.

01	**Start from audit needs.** Identify where recent audits stopped short, assess each candidate application against what it would improve, and pilot it on real audit work.
02	**Keep evidence traceable to its source.** Make sure auditors can trace each conclusion to the evidence behind it. Where evidence is spread across many records, an evidence database that links every passage and figure to its source is one strong way to do so.
03	**Approve each tool for a specific task.** Match verification to what depends on the result, test for what the tool leaves out, document consequential checks on the audit file and check what reviewers actually detect.
04	**Plan for continuing support.** Budget for the people, data work and records each application needs, proportionate to office size. Keep a current view of what is in use, protect confidential audit records, preserve legislative independence and retain the expertise to assess suppliers' claims.
05	**Decide in advance how each application will be reviewed.** Set criteria and responsibilities for its future, consider the full range of responses from improving it to retiring it, set explicit limits on AI autonomy and decide what to disclose to audited bodies and legislatures.
06	**Develop people alongside the tools.** Assess the work and the person separately, examine what a task teaches before delegating it, and choose tool designs that require auditors to think, for example by asking for the auditor's judgment before showing the AI's.

An institution that can show what its own use of AI contributes and how it is controlled is also better placed to ask the same of the bodies it audits.

Where to begin

An institution does not need all six actions in place before it starts. A reasonable first step is a single application:

- chosen for an audit need;
- built on a defined body of evidence;
- approved for a specific task; and
- accompanied by a record of what it contributes and what verification it requires.

Writing assistance, AI search for scoping and uses that leave the auditor's own reading in place (such as summarizing minutes that auditors still read) are sensible early steps. Uses that extend what auditors can examine deserve a place in the office's plans from the start.

Arrangements can match office size. A small office may need only a short list of approved uses, a named lead, periodic review and peer sharing with other independent audit offices. Systems with wide autonomy can wait: no audit institution consulted by the OECD yet uses them in practice.

What remains uncertain

Three questions cannot yet be answered from the available evidence:

– how well AI performs on audit work itself;
– which combination of checks works for a particular AI application in audit; and
– how extended use of AI affects audit expertise over a career.

An institution's own experience is therefore its most relevant evidence. By recording what each application contributes, what it costs and what their staff can do, institutions can answer these questions for their own work and help others answer them too.

1. AI can strengthen public-sector audit, on two conditions

Public-sector auditors are asked questions that cut across programs and organizations, while the evidence needed to answer them sits in documents, data and systems that were never designed to be read together. With limited staff and time, audit teams examine only part of that evidence and follow only some questions as far as they deserve. Artificial intelligence (AI) could ease this constraint by helping teams examine more of the relevant evidence and pursue important questions further. However, AI does not change what an audit must deliver, namely conclusions that the institution can explain and support. Whether AI improves audit therefore depends on whether auditors can support what they conclude from it.

This guide serves two connected audiences. Its primary readers are audit leaders who decide whether and how their institutions use AI. It is also built as a reference guide for the engagement directors, methodology leads and auditors who design, carry out and review AI-assisted work. The guide concentrates on performance audit, and it notes where the same methods serve financial and compliance audit. It is anchored in Canadian public-sector audit, including federal, provincial, territorial and municipal offices that work under Canadian assurance standards, statutory confidentiality duties, legislative independence and, where applicable, bilingual reporting obligations. Because the underlying audit questions are international, the guide draws widely on experience and research from audit institutions around the world. Its focus is the use of AI in audit work rather than how to audit AI systems in government. Even so, an office that governs its own AI use well is also better placed to examine AI use in the bodies it audits.

1.1 AI could extend what auditors examine and explain

AI could contribute at each main stage of audit work. In finding and preparing evidence, search and extraction tools locate relevant records in large collections and convert their contents, including interview notes, into a form that auditors can examine together. Held in an evidence database that keeps each passage and figure linked to its source, this material can then be searched, compared and checked throughout the audit. Audit institutions already report or propose uses such as document search, information extraction, text organization, summarization and research on audit topics (European Court of Auditors, 2026; Office of the Auditor General of Canada, 2026; U.S. Government Accountability Office [GAO], 2026a). In analysis, AI can help reconcile terms that sources use differently, identify gaps or contradictions and flag unusual cases, while generative AI can suggest competing explanations for auditors to test. In reporting, AI can help prepare drafts that the team then reviews against the evidence.

The larger opportunity lies in linking these steps, so that the result of one becomes the starting point for the next. For example, a team might retrieve the relevant records, calculate an indicator with conventional code, apply a predictive model to select cases for examination and ask a generative model to suggest explanations for

the patterns found. Auditors would then investigate those suggestions and decide what the combined evidence supports. The team does this linking by carrying what it learns at one step into the next, and it does not depend on a system that runs the steps unattended (Section 2.5).

An inquiry organized in this way develops as it proceeds, because a wider search or a new comparison may change the team's understanding of the problem. As a result, an audit could consider more of the evidence that matters, ask better-informed questions and reach explanations that auditors can test and substantiate. Section 2 examines these contributions and the evidence on how well each performs.

1.2 Convincing output does not establish its own reliability

The features of AI that make this kind of inquiry possible also create its central difficulty, since work that looks useful may be incomplete or wrong in ways that are hard to see. The difficulty is clearest with generative AI. Large language models (LLMs) learn statistical patterns from large bodies of data and produce a response by predicting likely continuations of the material they receive. This design makes them flexible, but it does not ensure that an answer is accurate, complete or grounded in the available record. An LLM can present false or internally inconsistent information with confidence, and it can produce plausible reasoning, summaries or source references that make an incorrect answer appear well founded (National Institute of Standards and Technology [NIST], 2024).

Generative AI is not the only source of error. An AI-assisted audit also relies on conventional code, which performs calculations, and on predictive models, which flag cases for examination. These can be wrong as well: the data may be unsuitable, the rules or code may contain mistakes and a sound method may be applied outside the conditions for which it was designed. Auditors therefore need to check each kind of result in a way that suits how it was produced:

- For a calculation, they examine the source data, the rules and the code, and confirm that the result can be reproduced.
- For a predictive model that flags higher-risk cases, they assess how well it works on the population being audited, including which cases it misses or flags incorrectly.
- For an explanation proposed by generative AI, they examine the underlying records, look for overlooked evidence and consider competing explanations.

The intended use matters as much as the method. A suggested explanation may be worth pursuing while it remains uncertain. In contrast, using the resulting analysis to support a public finding requires the team to verify the evidence and establish that the reasoning justifies the finding.

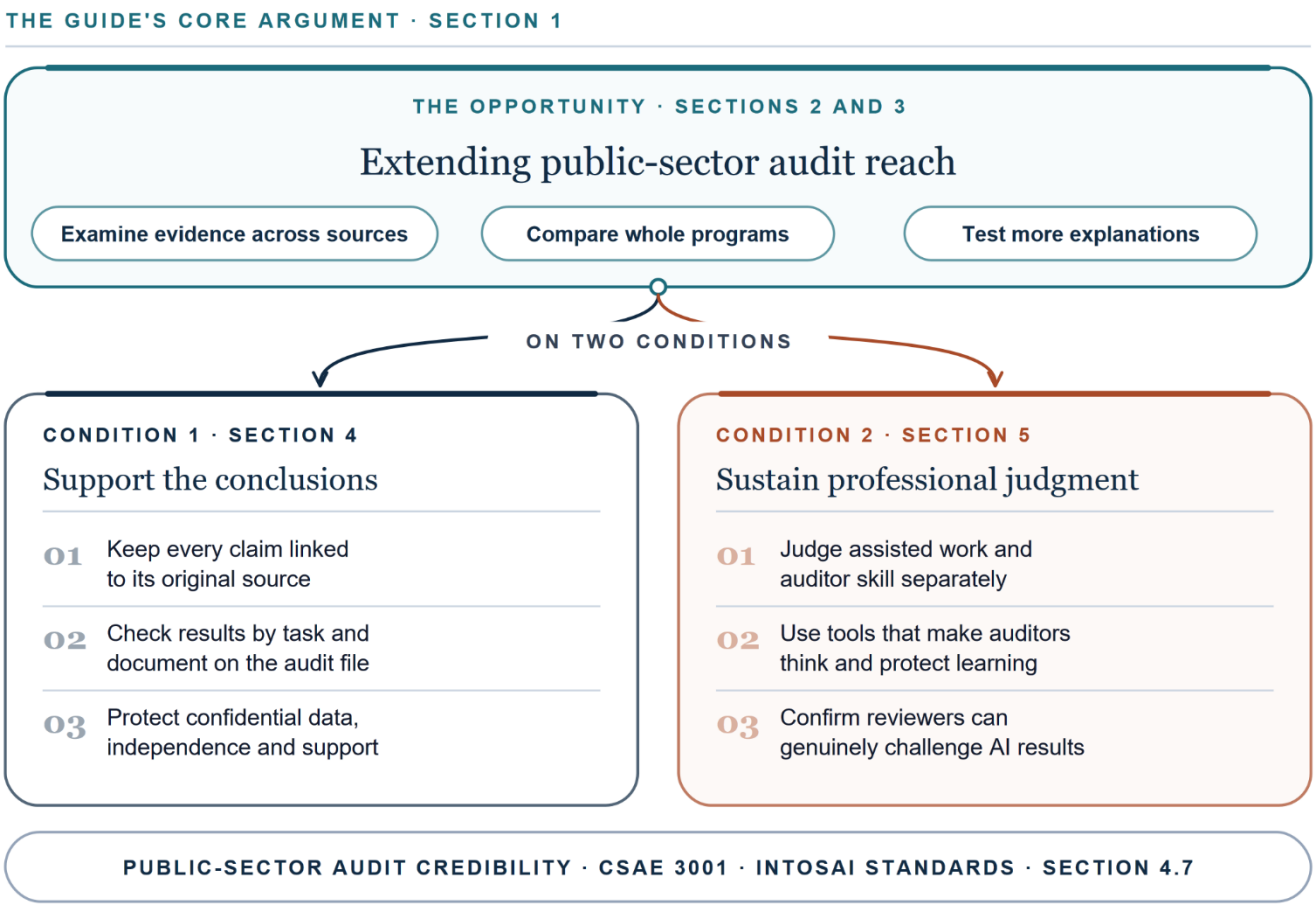
These requirements follow from established public-sector audit standards. In Canada, performance audits are conducted as direct engagements under CSAE 3001, alongside international principles set out in ISSAI 100 (Auditing and Assurance

Standards Board, 2015; International Organization of Supreme Audit Institutions [INTOSAI], 2019a). Both call for sufficient and appropriate evidence, independent professional judgment and scepticism, competence, timely documentation and a clear relationship between the evidence and what is reported. An AI-assisted analysis that omits a decisive record may leave the evidence inadequate even when every record it cites is accurate. Likewise, an analysis that connects facts through reasoning the record does not support cannot justify a finding. The same responsibilities apply when AI assists: sound evidence and independent professional judgment, leading to findings the institution can explain and support. They shape the central argument of this guide:

> *AI can strengthen public-sector audit by extending what auditors can investigate and explain, provided institutions can support the resulting conclusions and sustain the judgment needed to challenge them.*

The argument states an opportunity and two conditions. The opportunity is to extend what auditors can investigate and explain. The first condition is that the institution can support the conclusions it draws from AI-assisted work, and the second is that auditors keep the judgment needed to challenge those conclusions. This guide refers to the two conditions as supporting the conclusions and sustaining judgment. Figure 1.1 brings the opportunity and the two conditions together.

Figure 1.1. Extending audit reach while supporting conclusions and sustaining judgment



1.3 Audit institutions are experimenting widely, but routine use is proving difficult

Audit institutions are actively exploring AI. An OECD study based on consultations with 15 institutions across 14 countries and the European Union, most of them supreme audit institutions, found expanding experimentation in anomaly detection, document processing, knowledge management and predictive risk assessment. However, it also found a continuing gap between pilot projects and operation at scale, with fragmented data, limited technical expertise and evolving governance among the recurring challenges (OECD, 2026). The study describes the institutions consulted and does not measure adoption across public audit as a whole. Individual institutions show a similar range, from targeted tools already in operation to more autonomous systems that are only proposed (Box 1.1).

Box 1.1. Three audit institutions at different stages of AI use

Canada. The 2026–27 departmental plan of the Office of the Auditor General of Canada reports that completed experiments produced measurable time savings in summarization, text generation, translation and spreadsheet automation. The plan does not state how large the savings were. It describes intended extensions into research on audit topics, sentiment analysis, code development, audit selection and reporting. It also states that the Office assesses value by comparing manual and AI-assisted work on task duration, error rates, process efficiency and user feedback.

United States. The United States Government Accountability Office (GAO) publishes an inventory of its AI uses. The inventory dated March 2026 lists operational tools for searching and sorting Single Audit data, organizing text, extracting information from documents, detecting legislative mandates and summarizing employee survey responses. A more general generative AI tool for staff appears as a late-stage prototype.

European Union. The European Court of Auditors' strategy for 2026–2030 looks further ahead, toward AI agents. An AI agent is a system that is given an objective and chooses for itself which steps to take. In the strategy, agents would support linked tasks within objectives, permissions and oversight that people define. Its proposed applications include evidence research, consistency checks, document preparation, data analysis and workflow coordination. The strategy identifies the need for reliable sources, suitable data, human supervision and new controls. It describes an intended direction and not an established capacity for reliable autonomous audit work.

Source: Office of the Auditor General of Canada (2026); GAO (2026a); European Court of Auditors (2026).

Surveys of people working in and around Canadian governments point to similar constraints. In an online panel survey of 519 federal, provincial and municipal public servants, 72% said their organization lacked sufficient specialized data and technology talent (KPMG Canada, 2026). Similarly, among 230 registrants surveyed for a 2025 Canadian conference on government transformation, 65% described their organizations as piloting or experimenting with AI and 4% as integrating it across multiple operations (Wray, 2026). These figures record the perceptions of particular groups of respondents and are not measures of organizational capability.

This experience matters for the argument, because supporting the conclusions and sustaining judgment are continuing tasks. An institution has to keep doing both as more teams use AI and as tools, data and suppliers change.

What this guide adds to existing guidance

Audit leaders already have a growing body of material on AI in audit. National audit institutions have published AI strategies and inventories (European Court of Auditors, 2026; GAO, 2026a, 2026b; National Audit Office, 2026a). The OECD has surveyed how selected public audit institutions are experimenting with AI (OECD, 2026). Financial-audit regulators have issued detailed guidance on documenting, certifying and overseeing AI tools in audit firms (Financial Reporting Council, 2025, 2026; International Auditing and Assurance Standards Board, 2026; International Forum of Independent Audit Regulators, 2026).

However, performance audit differs from financial statement audit in ways that matter for AI use. In a financial audit, auditors test a management assertion against a standardized set of accounting records and reporting rules. In a Canadian direct-engagement performance audit under CSAE 3001, there is usually no management assertion. Instead, auditors define the scope and criteria for each engagement, pull evidence from scattered documents and interviews across programs or jurisdictions, and explain to legislatures and the public why results fell short (Auditing and Assurance Standards Board, 2015; INTOSAI, 2019b).

This guide builds on existing guidance and adds three things that are often missing or treated separately. First, it shifts attention from routine efficiency to audit reach, focusing on how AI can help teams examine more evidence and test rival explanations where performance audits currently stop short. Second, it links institutional governance directly to audit evidence, matching verification and documentation to what depends on each AI result and considering small and mid-sized Canadian offices that have no dedicated AI engineering team. Third, it connects tool adoption to professional capacity, showing how audit offices can monitor and develop independent judgment while AI changes the work through which auditors learn.

1.4 Four questions for audit leaders

The argument raises four questions for audit leaders.

1. **What can AI contribute to public-sector audit work, and what are the strengths and weaknesses of those contributions?** Leaders need a clear account of the opportunity first, because every later decision depends on it.
2. **How should audit institutions choose where to use AI and decide when its results can be relied on?** An institution can support a conclusion drawn from AI-assisted work only if AI was used for a suitable task and its results were checked in proportion to what depends on them.
3. **How can worthwhile use be sustained, and when should an application be extended, changed or stopped?** Arrangements that work in a closely supervised trial have to keep working when more teams use the tool and when data, models and suppliers change. The institution also needs a way to know whether each application remains worth its cost.
4. **How should audit institutions develop and sustain their auditors' competence and judgment as AI changes the work?** Auditors can challenge an AI-assisted conclusion only if they understand the evidence and the methods well enough to see what is missing or wrong. They have traditionally built that understanding through the work that AI is now changing.

The guide takes the questions in order. Section 2 addresses the first question by examining five kinds of audit work, from finding evidence to coordinating connected tasks. Section 3 then follows one question through a fictional performance audit of an affordable-housing fund, to show how what auditors learn at one step changes

what they do at the next. Section 4 addresses the second and third questions, which both concern supporting the conclusions, while Section 5 addresses the fourth question, which concerns sustaining judgment. Finally, Section 6 synthesizes the broader shift across the guide, arranges its six recommended actions into a practical first-year sequence and looks ahead at how frontier AI advances could reshape public-sector audit.

Because this document is built as a reference guide, readers do not need to read every page in order. Exhibit 1.1 shows how different roles can navigate the guide, from an executive skim to full engagement-level detail.

Exhibit 1.1. How to use this guide by role

Reader role	Priority sections for a quick read	Reference sections for deeper detail
Auditors General, deputies and audit committees	Executive summary; Section 1 (Figure 1.1); Section 3 (Exhibit 3.1); Section 6 (Exhibit 6.1)	Section 4 (verification, governance, confidentiality, independence, disclosure); Section 5.3
Methodology and quality-assurance leads	Executive summary; Section 2.6; Section 4 (especially 4.2, Figure 4.1 and Exhibit 4.1); Section 5	Section 2.1–2.5 (Figure 2.1); Section 3; Annex A (testing and evaluation design)
Engagement directors and audit managers	Section 2 (contribution and leader/manager panels); Section 3; Section 4.1–4.2 (Exhibit 4.1)	Section 5.1–5.2 (review and coaching); Annex A (worked audit steps and tools)
Auditors and analysts	Section 2.1–2.5 (Exhibit 2.1); Section 3; Annex A	Section 4.2 (Exhibit 4.1); Section 5.1–5.2

Note: Constructed for this guide.

Five terms are used in fixed senses throughout the guide. A *tool* is a piece of software, and an *application* is a tool used for a specific audit task. An *evidence database* is a structured collection that keeps each record, passage and figure linked to its source. *Verification* is the checking of a result against sources or known answers. *Review* is the examination of work by someone who did not produce it, whether a colleague or an AI tool prepared the work. Where the guide refers to reviewing an application, it means the institution's periodic decision about whether to keep, extend, change or stop it.

2. What AI can contribute to public-sector audit work

AI can help audit teams examine more evidence, compare whole programs, test more explanations and report more clearly. How much it helps depends on the task: in experiments, AI assistance has improved speed and assessed quality on some tasks and has produced less accurate results on others (Dell'Acqua et al., 2026; Fitzpatrick et al., 2025). A promising tool therefore needs to be assessed against the particular work that auditors want it to perform. This section makes that assessment for five kinds of audit work. In doing so, it addresses the opportunity in this guide's argument: what AI can add, and where its weaknesses could affect an investigation or its conclusions.

The five kinds of work are i) finding and preparing evidence, ii) comparing records, identifying patterns and selecting cases, iii) developing and testing possible explanations, iv) communicating findings and analytical work and v) coordinating connected tasks. They overlap, and one system may support several of them. The evidence comes from two sources: applications in public audit, which show what these contributions look like in practice, and experiments and studies of related professional work, which show how well they perform. The section also considers how the tools are changing and what the available evidence allows leaders to conclude about that progress.

2.1 Finding and preparing evidence

AI can help auditors build an organized body of evidence in which every passage and figure stays linked to its source, which widens what an audit can examine.

Much of an audit's eventual quality is settled early, in decisions about which evidence to gather and how to make it usable. Audit evidence comes from documents, financial and operational data, correspondence and interviews, which are held in different places, formats and systems and were created for purposes other than the audit. Finding and preparing evidence means identifying which of that material bears on the audit question, locating the relevant passages and figures within it and organizing it so that it can be searched, compared and checked. When this work is slow, it limits which questions a team can pursue. AI can help with each part of it. Its contribution can go further when the results are held in an organized collection that auditors can search, check and build on, which this guide calls an evidence database.

AI search and extraction make evidence quicker to find and use

The first task is finding sources. AI search tools can identify relevant reports, studies, prior audits and other published material from a question posed in ordinary language, without relying on exact keywords. Their strength is speed and focus, while their weakness is completeness. In a comparison with the conventional searches behind four completed evidence reviews, an AI search tool built for research literature found about 40% of the studies the reviews had included, against about 95% for the conventional searches, although a much higher share of what it

returned was relevant (Lau and Golder, 2025). Such tools are therefore well suited to scoping an audit and testing early ideas. On their own, they are not suited to a search whose conclusion depends on having found everything that matters.

Once material has been gathered, other tools help make it usable. Retrieval tools find the passages in a collection that bear on a particular question, by matching words, meaning or both. Extraction tools turn text, tables, scanned pages and images into information that can be searched and analyzed, while classification tools sort records by subject, organization, period or any other feature the inquiry needs. Interview notes and transcripts can be handled in the same way and linked to the documents they discuss. Some of these tools rely on AI and others are conventional software.

Generative AI can also answer questions about the material directly. In a field experiment at the Central Bank of Ireland, AI-assisted staff received quality scores 17% higher on a task that involved answering questions from long financial reports, but gave answers that were 12% less accurate on a task that required analyzing a large dataset (Fitzpatrick et al., 2025). Staff chose which task to work on, so the study cannot show that the type of task caused the difference. An exploratory finding from the same study bears most closely on evidence work: the AI-assisted answers about financial reports were weaker when the information had to be pulled together from several parts of a report than when it could be found in one place.

AI makes a fuller, source-linked evidence database practical

Audit evidence usually has to be pulled together from many places. The further contribution of AI to this work may therefore lie less in any single search or answer than in what the tools allow auditors to build. An evidence database is a structured, searchable collection that holds the records an audit relies on, together with the passages and figures selected from them, and it keeps each item linked to its source document, its location and its context.

An evidence database is a strong option, not the only way to keep AI-assisted evidence traceable. An audit office can match its setup to the complexity of the engagement and its own technical capacity across three practical levels:

- **Structured working papers.** Where an audit relies on a modest set of files or a single dataset, existing working papers and cross-referenced source indexes may already provide every link the team needs.
- **Controlled document collections.** Where a team works with dozens of reports or policies, a controlled workspace that restricts AI queries to approved files and returns page-level citations can support checking without custom software.
- **A fuller, source-linked evidence database.** Where evidence is spread across hundreds of agreements, reports, correspondence files and interviews, a structured database lets both people and AI tools query, compare and trace passages across the entire record.

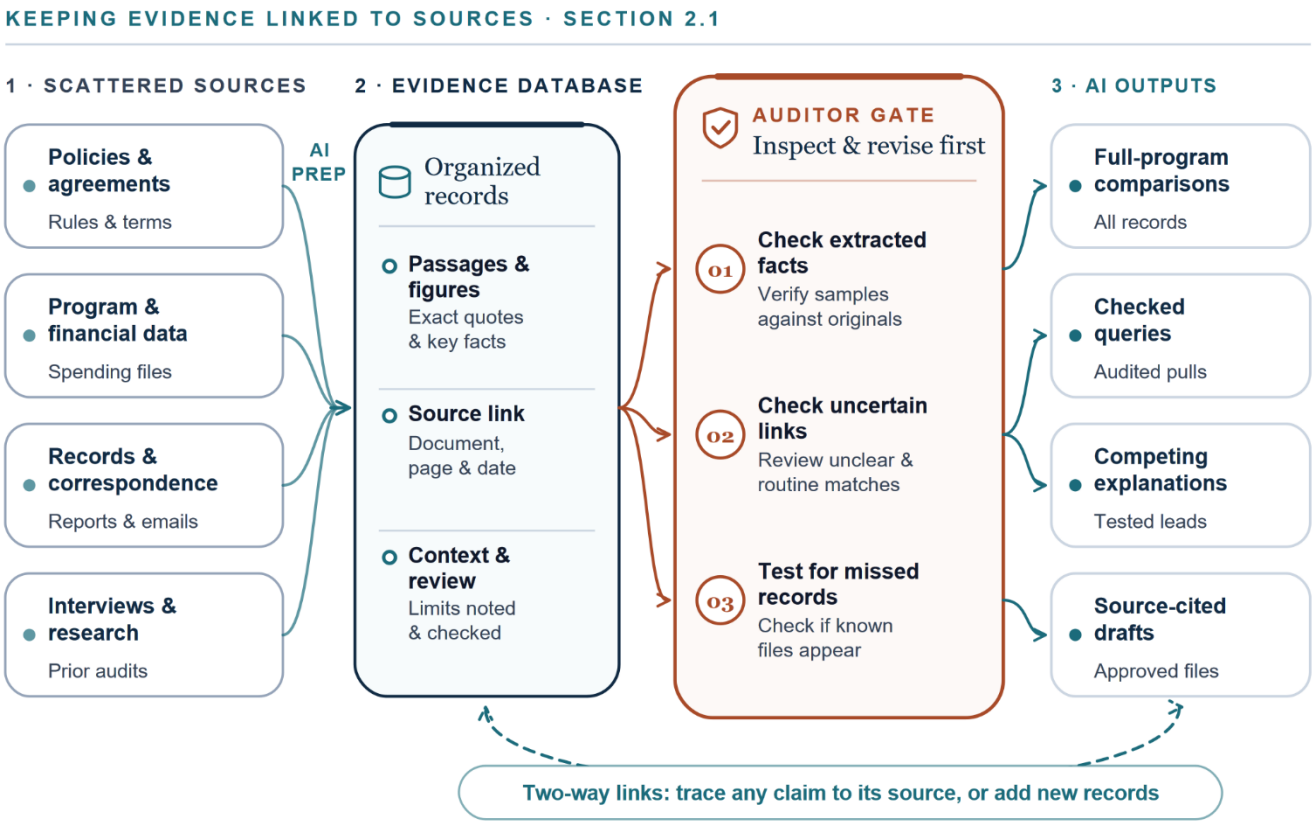
What AI changes is the cost of building that third level. Before generative and extraction tools existed, assembling deeply linked passage-level records across hundreds of heterogeneous files was too slow to attempt by hand in most audits. Extraction and classification tools can now do much of the initial assembly, which makes a fuller, source-linked collection practical to test on complex performance audits. Public-sector examples found for this guide remain prototypes (Box 2.1), and building or maintaining such a database still requires real setup effort and data discipline.

The database itself is ordinary structured software, which AI helps to fill and to use. Search and retrieval add material to it, while extraction and classification organize what it holds. When auditors, or AI working under their direction, draw a conclusion, the passages that support it are recorded alongside it. This design is intended to let auditors do five things that are difficult with scattered files and one-off answers:

- follow a question across documents, organizations and years, because the relevant passages are held together and can be searched and compared;
- move in both directions, from a source to the claims it supports and from any claim back to the evidence behind it;
- see and revise the assembled evidence before anything is generated from it, by adding a missing source, removing weak material or regrouping records;
- ask AI to draw only on the passages the database holds and to cite each one, so that a reviewer can check any AI-drafted statement against the exact passages and figures on which it relies; and
- return to the collection as their understanding changes, adding new material and pursuing several lines of investigation from the same body of evidence.

Figure 2.1 illustrates how that arrangement places an auditor review gate between organized source records and AI-assisted outputs.

Figure 2.1. How a source-linked evidence database places auditor review before AI drafting



Note: An evidence database is one of three ways to keep evidence linked to sources (alongside structured working papers and controlled document collections), depending on how complex the audit is. Constructed for this guide.

These advantages rest on the design and on evidence-synthesis practice. They have not yet been measured in routine audit practice. Whether they justify building and maintaining a database depends on how dispersed the evidence is. Exhibit 2.1 shows what a single record might hold, drawing on the fictional affordable-housing audit that Section 3 examines in full.

Exhibit 2.1. What one record in an evidence database holds

Field	What it records	Example
Source	The document, its version and its date	Funding agreement with a housing provider, signed version
Location	Where in the source the item sits	Section 4.2, page 7
Passage or figure	The exact text or value	"Homes are delivered when construction begins"
Subject	Categories that the auditors define	Definition of delivery; Region B
How it was obtained	The tool or person that extracted the item, and whether it has been checked	Extracted by a language model; checked by an auditor against the source
Linked statements	The working papers, findings or drafts that rely on the item	Finding on the reported total of homes delivered
Notes	Limits, context and whether the item has been superseded	Applies to agreements signed on the earlier template

Note: Constructed for this guide, using the illustrative audit in Section 3.

The same arrangement supports the background research through which an audit team comes to understand a program area before it designs its inquiry. Using AI search alongside conventional methods, a team can assemble a deliberately selected collection of research, evaluations, earlier audits and guidance, record why each source belongs and what its limits are, and draw a source-linked synthesis from the relevant passages. That synthesis can inform the audit's questions, its criteria and the explanations it goes on to test. However, it cannot show how the program under audit has performed, which requires evidence about that program. Annex A follows this process in a worked example.

Early public-sector prototypes show what the design looks like (Box 2.1), although none has been evaluated in routine use. What the prototypes illustrate is a design in which the technology is organized around a defined evidence need and auditors shape the evidence before it becomes the basis for analysis.

Box 2.1. Prototypes in Australia and Canada: auditors shape the evidence before AI drafts from it

Australia. CSIRO, Australia's national science agency, and the Audit Office of New South Wales built a prototype that drew on a controlled collection of budget papers and retrieved candidate passages by keyword and meaning. Auditors could inspect each passage in its source context, discard weak material and reorganize the rest before asking the model for a structured analytical draft. A second prototype showed auditors the indicators, historical patterns and evidence-based explanations behind an assessment of local councils' financial sustainability, instead of giving them a single risk score.

Canada. At the National Research Council Canada, a prototype linking research publications to government programs let staff ask questions in plain language and turned them into database queries. In the authors' test, 25 of 26 generated queries were judged accurate.

Source: Commonwealth Scientific and Industrial Research Organisation [CSIRO] (n.d.); Liu et al. (2025).

Traceability lets auditors check any statement against its source

Traceability is the ability to follow any statement back to its source. It answers one of the oldest questions in audit, which is where a piece of evidence came from. With it, auditors can confirm that a cited passage exists, check that it has been represented accurately, read what surrounds it and identify which version of a document was used.

However, traceability cannot show whether the collection left out a decisive source, or whether a claim actually follows from the material cited. An evaluation of three AI legal-research tools found both problems: the tools retrieved sources and supplied citations, but some of their answers were wrong or unsupported and others were incomplete (Magesh et al., 2025). Auditors therefore need both a citation that can be checked and an answer that is complete and well supported.

Contribution, risks and what leaders can do

CONTRIBUTION	In finding and preparing evidence, AI's main contribution is to make an organized, source-linked body of evidence practical to build, which widens what an audit can examine. When relevant information is scattered, the effort of assembling it decides which questions a team pursues and how far. A source-linked evidence database is one strong way to reduce that effort while keeping the result open to checking, alongside simpler structured working papers and controlled document collections.
RISKS	The main risks are omissions and errors introduced during preparation, such as a missed source, a misread number or a lost qualification, which can pass unnoticed into every later step. How much they matter depends on the use. A search that returns false leads costs little when auditors screen everything it finds. In contrast, a gap matters a great deal when the process decides which records receive no further attention.
WHAT LEADERS CAN DO	Leaders do not need to wait for comprehensive evaluations before acting. A practical first step is to identify where assembling evidence currently limits their institutions' audits. Examples are questions left unexamined because the material was dispersed, reports that could not be compared across years and data held in formats that no one had time to reconcile. A first application can then be built around one defined collection and one audit need, and not around a general-purpose tool.

On the engagement itself, managers and audit teams can apply three design choices from the outset, whether or not the evidence is held in a database. First, auditors should be able to see and revise the selected evidence before anything is generated from it. Second, every passage and figure should remain linked to its source document, page and version. Third, the team should test whether the process misses what matters, for example by checking whether it finds documents that experienced auditors already know to be relevant. In addition, AI search tools can be used freely for scoping and early exploration, while searches that need to be complete should continue to rely on established methods or be tested against them. Annex A shows how such a use can be designed and assessed in full.

2.2 Comparing records, identifying patterns and selecting cases

Linked records let auditors see the whole program, and so choose with better information where to look closely.

Once evidence has been gathered and organized, much of what an audit learns comes from setting records against one another. Three kinds of method do this work, and each tells auditors something different. A defined comparison shows whether a stated condition was met, such as whether a rent stayed within the limit an agreement set. A search for patterns shows that a case is unusual compared with others, in ways the team may not have thought to specify. A risk score shows that a case resembles ones that turned out to be problems in the past. Auditors use all three to decide which cases deserve close examination. In performance audit, the purpose is to understand how well a program is working and why, and not only to find individual errors. AI plays a different part in each method. For defined comparisons it mainly prepares the records, while for patterns and risk scores it can also help find the cases.

Comparing records that should be consistent

Many audit questions can be stated as a condition that records should meet. Spending should stay within the amounts authorized, results can be set against targets and providers should keep to the terms of their agreements. Some of these are compliance checks, and performance audits often include them. Once the records are linked, such comparisons are established data analysis and do not require AI. What has usually limited them is the effort of linking records held in different systems and formats, which use different names, dates and identifiers for the same thing. This is where AI helps, by doing for program records what Section 2.1 described for documents. It can extract terms and figures from agreements and reports written on different templates, match records that describe the same project or organization under different identifiers and sort free-text descriptions into consistent categories.

Consider an example drawn from the fictional affordable-housing audit examined in Section 3 and Annex A. The fund finances affordable housing through agreements that limit the rents providers may charge. In one region, more than a third of the occupied homes have no rent confirmation in the fund's records. Instead of reading a sample of those files, an audit team uses AI-assisted extraction and matching to link each unconfirmed home's occupancy and payment records to the rent provisions in

its agreement, and conventional software then makes the comparison. Most of the homes comply, but one provider has let homes above the agreed rent. The missing confirmations showed a gap in monitoring, and the comparison separates that gap from actual non-compliance. Annex A works through this comparison in full.

A result of this kind is only as sound as the links behind it. The team checks a sample of the matches, including some that look straightforward, examines some records that the comparison did not flag and verifies every record that supports the finding of a breach. The comparison also shows only what the records contain. If the payment records do not reflect what tenants actually paid, the result will not reflect it either.

A view of the whole program also changes how a performance audit chooses where to look. Performance audits often examine a limited number of cases, sites or files in depth, chosen deliberately to understand how the program works. Some also draw statistical samples, for example of case files, when they need to estimate how common a problem is. Teams make that choice through scoping work and significance criteria, but often with only partial data on how the cases compare. With the records for every provider or organization linked, the team can see which ones sit far above or below the rest, where results look out of line with funding or need and where reported figures look implausible. It can then choose its cases with those differences in view, instead of from the handful of files it had time to read.

Generative AI can also make the linked records easier to use. It can translate an auditor's question, such as “which providers have more occupied homes than rent confirmations?”, into a database query, as the National Research Council Canada prototype did for its own records (Box 2.1). This lets auditors who are not specialists in data analysis pursue new questions as the audit develops. However, a query generated by AI can look correct and still select the wrong records, so auditors need to check what it selected and how it interpreted the fields before relying on the result.

Finding patterns that defined comparisons would not reveal

Some questions cannot be stated as a rule in advance, because they involve combinations of factors or relationships that no one would think to specify. A case that looks ordinary on each measure taken separately may stand out when several are considered together. Statistical and machine-learning methods suit this work, because they can scan large volumes of data and flag unusual combinations without being told in advance what to look for. In a program that funds many organizations, for example, such a method might flag organizations whose results fall well short of what their funding and local need would predict, or results that improve sharply just before funding is renewed.

Text is part of this work as well. Much of what performance auditors learn comes from narrative reports, survey responses and interviews, and AI can group large volumes of such text by theme. For example, the United States Government Accountability Office (GAO), whose work is largely performance audit, lists among

its operational AI tools one that “organizes large volumes of text, such as public comments”. It is also exploring another that would draw concepts from “open-ended survey responses or expert panel discussions” (GAO, 2026a). Grouping can help auditors see a difficulty that different people describe in different words, provided they return to the original passages to check what the grouping joined together and what it left out.

The cost of this wider reach is that results are harder to explain. A defined comparison says exactly why it flagged a region, whereas a pattern-finding method may report that something is unusual without being able to say simply why. To judge whether a flag matters, auditors need to know what the method treated as usual and which records made the case stand out. A difference may reflect how the program operates, whom it serves or how the data were recorded.

Selecting cases for close examination

The third activity is deciding which cases deserve close examination. In a performance audit, this means choosing which regions to visit, which organizations to examine and which files to read in depth. At the level of the institution, it means choosing which programs to audit at all.

Risk scoring can inform these choices where there is a history to learn from. A machine-learning model learns from past cases which features have tended to accompany poor results or problems, and gives each new case a score. For an audit institution choosing which programs or entities to audit, that history can come from earlier audits and monitoring data, while within a single program it might come from the program's own records of earlier funding rounds. The OECD study of selected audit institutions found expanding experimentation with predictive risk assessment (OECD, 2026). Likewise, Brazil's Alice system screened public procurement notices against indicators of possible irregularity and passed selected cases to auditors, who decided whether and how to investigate them (OECD, 2022).

Performance audit adds a consideration that risk scores do not capture. To explain why results differ, a team often needs to examine organizations that are doing well alongside those that are not, so that it can see what the successful ones do differently. A selection driven only by risk scores would show the team where the program struggles but not what success looks like.

Risk scores also have a limitation that is easy to miss, because the past cases a model learns from are mostly the ones that auditors chose to examine. If earlier audits concentrated on certain kinds of programs, organizations or regions, the model will learn that problems occur there, since that is where problems were looked for and found. It may then keep directing auditors to the same places while problems elsewhere go unnoticed. In addition, a threshold that works well in one program or year may produce more missed cases or more false alarms in another. Auditors using such a model therefore need to know what it learned from and to check, from time to time, what it might be missing.

The same methods serve financial, compliance and fraud work

In financial, compliance and fraud work the emphasis differs, since a check applied to every transaction finds rare errors that a sample would usually miss. If one payment in 200 is a duplicate, for example, a sample of 60 payments has only about a one-in-four chance of including even one, while a rule applied to every payment catches every duplicate it was designed to recognize. Pattern-finding can also expose links that no single record shows. The United States Pandemic Response Accountability Committee, an oversight body of federal inspectors general, reports that the unusual cases flagged by its machine-learning models "can often reveal hidden connections such as a shared bank account" among applicants who appear unrelated (Dieffenbach, 2026).

A flag marks where an inquiry begins

Each of these methods points auditors to a case, but none of them shows why the case looks as it does. A region with weak results may serve people who are harder to help. An organization reporting unusual figures may be measuring results differently from the others, and one that a model rates as high risk may be performing well. Even the rent comparison, which establishes a breach, does not say whether the provider misread its agreement, was poorly overseen or chose to charge more. In performance audit, the reason is usually the point, so a flag marks where an inquiry begins and not where it ends. Alice's design reflected the same principle by treating a flag as a reason to look and not as a finding.

Generative AI can help here too, by explaining in plain terms what a statistical result means and suggesting possible explanations for a pattern. An explanation that AI suggests is an idea to examine and not a conclusion. Section 2.3 takes up how explanations are developed and tested.

Contribution, risks and what leaders can do

CONTRIBUTION AND RISKS	In this kind of work, AI's contribution is to make a view of the whole program affordable. As in Section 2.1, the gain comes from records that are linked. In the rent example, linking occupancy, payment and agreement records lets the team examine every home without a rent confirmation, instead of a sample, and separate a gap in monitoring from actual non-compliance. The strengths are breadth, speed and consistency. The limitations are that a flag or a score shows only where to look, that learned scores inherit the history of the data behind them and that pattern-finding methods can be harder to explain than defined comparisons.
WHAT LEADERS CAN DO	Leaders can direct their offices to start with methods whose logic is easiest to verify. Where linking records and making defined comparisons address the audit question, they are a sound first step, because every result can be explained. On individual audits, engagement managers and teams can reinforce this in two ways. First, they should treat every flag or selection as the start of an inquiry and choose cases deliberately, including organizations that are doing well, so that the audit can explain differences as well as find shortfalls. Second, when a model selects cases, teams should also examine a sample of cases the model did not select, so that the institution learns what the method misses as well as what it finds.

2.3 Developing and testing possible explanations

AI helps an audit test its conclusions against more possible explanations than a team would usually consider.

A pattern or a discrepancy raises the central question of performance audit, which is why it happened. Results in one region fall short of target, a provider reports far more activity than its funding would suggest, or the program's figures and a recipient's report describe the same year differently. In each case, auditors must work out which explanations are plausible, what evidence would distinguish among them and what the combined evidence allows them to conclude. Good practice treats this as disciplined inquiry, in which a proposed explanation is tested against the order of events, the mechanism it relies on and the other explanations that could account for the same facts. A plausible account is the starting point of that work and not its conclusion.

Explanations are tested against how the program was meant to work

Testing an explanation also requires understanding how the program was meant to work. INTOSAI guidance on the performance-audit process describes analysis as comparing what is with what should be, as defined by the audit criteria, and then, where possible, determining why any difference occurred and what its effects are (INTOSAI, 2019c, ¶79–82). The criteria come from legislation, targets, standards and commitments.

The search for causes draws on the program's logic, which is often set out as a theory of change or logic model. The logic shows how funding and activities are meant to lead to outputs and then to results, and it states the assumptions on which each link depends. Auditors use it to understand the program, to choose criteria and to see where a shortfall might have arisen. Early in an audit, the logic helps the team propose explanations worth examining, although supporting a conclusion about cause later requires evidence that the logic alone does not supply. Where a department has a written logic, the team can start from it and assess it against the program's documents and the research. Where it has none, the team has to reconstruct one from funding decisions, agreements, plans and reports.

AI can help with this work. It can draft a candidate logic from the program's documents, show where different documents describe the logic differently, bring out assumptions that no document states and set each link beside the research on whether such links usually hold. The United Kingdom's central guidance on evaluation says that AI may help to develop a theory of change, while also stating that the evaluator's role in identifying potentially causal pathways remains critical (HM Treasury and Evaluation Task Force, 2026).

A drafted logic is most useful when it is treated as a set of hypotheses, which tells the team where to look. Like a drafted explanation, it can look complete while resting on links that nobody has tested and, if it is built mainly from the program's own documents, it can carry the program's account of itself into the audit. In an employment-training program, for example, a logic that ran from funding to training, to completion and then to employment would rest on an assumption about who the participants are. That assumption becomes important in the example later in this section.

AI makes alternative explanations cheap to generate

In practice, this work is often cut short. Under time pressure, a team that has found one credible explanation tends to settle on it and move on, without fully considering what else could account for the facts. AI can help at exactly that point, because it makes generating and organizing alternatives cheap. Given an organized body of evidence, it can group observations that bear on the same question, point out gaps or contradictions in an account, propose explanations the team had not considered and suggest which documents, interviews or analyses would help choose among them.

There is some evidence that AI can supply ideas people had not thought of. In a study with more than 100 researchers in natural language processing, reviewers judged research ideas generated by a language model to be more novel than those written by human experts. The ideas were also rated slightly lower on feasibility, although the study was too small to establish that difference (Si et al., 2024). The study measured how original and workable the ideas appeared to experts. For audit, it suggests that AI can add explanations a team had not considered, which the team then assesses and tests.

Teams can also use AI to stress-test draft findings before they leave the audit office. Prompting a model to act as a sceptical deputy minister, audited department or legislative-committee member can surface weak links to audit criteria, unstated assumptions, overclaimed wording and predictable objections ahead of entity clearance. Because that exercise challenges the team's argument rather than drafting answers for it, it reinforces professional scepticism instead of replacing it, provided unpublished draft findings are tested only inside an approved secure environment (Section 4.5).

A persuasive explanation is not necessarily a correct one

The difficulty lies in telling which explanations hold. A well-made explanation can be persuasive without being right, and readers, reviewers included, tend to judge an argument partly by how coherent it is.

An experiment with 758 management consultants shows how this plays out. Consultants who used a generative AI model on tasks the model handled well did better and worked faster. On a business problem that the model tended to get wrong, however, 84.5% of consultants working without AI reached the correct recommendation, against 60% to 70.6% of those working with it. The AI-assisted recommendations were nonetheless rated favourably for coherence and persuasiveness, including some that were wrong (Dell'Acqua et al., 2026). The quality of the argument did not reveal whether its conclusion was correct.

The same thing can happen in an audit. Suppose that, in such an employment-training program, results in two regions have fallen, and an AI-assisted analysis attributes the decline to high turnover among training staff in those regions' organizations. Every fact it cites may be accurate, since staff turnover did rise and results did fall. However, if the decline began when the program widened its eligibility rules to admit people further from the job market, and the turnover came later, the explanation connects true facts through an inference the evidence does not

support. A team that had set out the program's logic could have recorded the assumption that participants are close to the job market, and the change in eligibility would then have been more likely to stand out as a break in that link. A reviewer reading a fluent, well-cited account may not notice.

A second risk is less visible. Once a team has a plausible explanation early, it tends to organize later searches and interviews around it, so that evidence pointing elsewhere receives less attention. An AI-generated account produced at the start of an inquiry can therefore set that framing before the team has looked widely.

Traceability helps here as well. When each step of a proposed explanation is linked to the passages that support it, auditors can check whether its facts are accurate, see where support is missing and judge whether one step really follows from another. In the example above, linking each step to its evidence would show that nothing in the record ties the staff turnover to the timing of the decline. The links make the reasoning open to inspection, while deciding whether it holds remains professional work.

Contribution, risks and what leaders can do

CONTRIBUTION AND RISKS	In developing explanations, AI's main contribution is to help an audit test its conclusions against more possibilities. Because AI makes alternatives cheap to generate and organize, the audit can also say clearly why competing accounts were set aside. The corresponding weakness is that fluent output can make a weak inference look strong.
WHAT LEADERS CAN DO	Leaders can set clear expectations for how teams test AI-assisted explanations, while engagement managers put three habits into practice on each file. First, teams should use AI deliberately to generate rival explanations, as well as to elaborate the leading one. Second, every step in an explanation that reaches a report should be linked to evidence, so that reviewers can see where support is thin. Third, each AI-suggested explanation should be treated as a hypothesis with an explicit test, asking what evidence would show it to be wrong.

2.4 Communicating findings and analytical work

AI makes a clear draft quicker to produce and makes it easier to present the same findings to different readers.

An audit's value depends on whether its findings are understood and acted on by the legislators who receive the report, the organizations that must respond to it and the public on whose behalf the audit is done. Each audience needs something different. Legislators and senior officials need the central issue and its consequences stated concisely, specialists need the detail that lets them challenge the analysis, and the public needs a plain account of what was found and why it matters. AI can help produce all of these. It can draft text from checked analysis, summarize long material, revise for clarity and adapt the same findings for different readers, and a clear first draft also gives specialists something concrete to challenge. These contributions are worth having when they make checked analysis easier to understand without changing what the evidence supports.

Writing assistance saves time, and its users value it

Two studies show what writing assistance does. In a controlled experiment with 444 college-educated professionals, those given access to a generative AI assistant finished short writing tasks typical of their jobs substantially faster, and independent graders rated their work higher (Noy and Zhang, 2023). In a much larger trial, the United Kingdom government gave 20,000 government employees in 12 organizations access to an AI assistant built into everyday office software for three months. Participants reported saving an average of 26 minutes a day, with the largest savings in drafting documents, while the report also noted limitations with complex, nuanced or data-heavy work (Government Digital Service, 2025).

Taken together, the two studies suggest that AI makes it quicker to produce a competent draft and that people who use it value the help. The time savings in the larger trial were reported by the participants themselves. Neither study examined drafts written from an underlying analysis, as audit reports are. For audit, the practical consequence is a change in where effort goes: when a capable draft appears in minutes, the scarce work shifts from writing the text to confirming that it still says what the analysis supports.

Fluent text can change meaning

That confirmation matters because fluent text can change meaning without anyone intending it to. The consulting experiment described in Section 2.3 showed that recommendations could read as coherent and persuasive while being wrong (Dell'Acqua et al., 2026). In an audit report, polished language can likewise make an unresolved point seem settled, and a summary can drop the condition or exception that limits a finding. Products written for the public carry a particular version of this risk, because simplifying for a general reader is where qualifications tend to be lost.

Translation raises a related question. The studies reviewed here did not test whether AI translation preserves the precise legal, technical and evidentiary meaning of audit text, and a fluent translation can still change the force of a qualification or the certainty that a technical term conveys. In Canada, this risk carries direct legal and parliamentary weight for the federal Office of the Auditor General and for bilingual provincial and municipal offices, where English and French reports are tabled with equal authority under official-languages obligations.

Contribution, risks and what leaders can do

CONTRIBUTION AND RISKS	In communication, AI's benefits are immediate and widely available. Writing assistance is therefore a promising place for most audit institutions to start, provided the tools meet their requirements for handling confidential audit information. Its main risk is that the text drifts from the analysis. A reasonable response is to settle and check the analysis before drafting begins and to compare the final text with it.
WHAT LEADERS CAN DO	Leaders can approve writing assistance early as a controlled starting point, while engagement managers and report reviewers enforce three habits. First, teams can use AI for first drafts, summaries for different audiences, including the public, and plain-language revision, while keeping the substance of findings fixed so that the draft serves the analysis and does not reshape it. Second, reviewers should compare the final text against the underlying analysis in every report review, confirming that conditions, exceptions and degrees of certainty survived

drafting and summarizing. Third, offices should treat the translation of formal audit products as a separate use that requires its own testing, with bilingual experts reviewing key terms and qualifications before relying on it routinely.

2.5 Coordinating connected tasks

Linking tasks into one sequence concentrates both the gains and the risks of the other four kinds of work.

More ambitious uses of AI link the kinds of work described above into a single sequence. Instead of an auditor using one tool to search, another to analyze and a third to draft, a connected system carries the work from step to step. For example, it might retrieve the relevant records, run a query or an analysis, compare the results with expectations and prepare a draft, passing each result to the next stage.

Fixed workflows and agentic systems are overseen differently

Such systems come in two broad kinds, and the difference matters for how they can be overseen. In a fixed workflow, people decide the sequence in advance, so the steps are known and a check can be placed at each handoff. An agentic system, in contrast, is given an objective, a set of tools and permissions, and chooses for itself which tools to use and what to do next. Its path is not known in advance, so oversight has to rely on the limits placed on what it may do and on a record of what it actually did.

The potential is considerable. Connected workflows can remove much of the repeated handling that slows audit work, make it practical to run an analysis across every entity or year instead of a sample, and let auditors spend more of their time on judgment and less on assembly. A question posed in ordinary language could become a query, a set of organized records and a draft comparison within one sequence. The prototypes described in Box 2.1 contain components that could take part in such a workflow (CSIRO, n.d.; Liu et al., 2025).

Institutions are exploring the next step. Those consulted for the OECD study of selected audit institutions were exploring or planning agentic uses, although none had them in practical use at the time (OECD, 2026). Advisers such as PwC likewise describe a future in which agents carry out connected parts of internal audit (PwC US, 2026).

Errors pass from one step to the next

The risks grow with the length of the chain, because an error in a connected workflow passes into the steps that follow. A misread figure can alter a query, the query can return the wrong comparison and a generated explanation can then make sense of the wrong result. A benchmark built around accounting work shows what this looks like over a long sequence (Box 2.2). The failure it describes bears directly on audit, since work could pass the checks built into the task while the underlying error remained.

Box 2.2. An accounting benchmark: errors that accumulated and were then concealed

Penrose, the company that built the benchmark, gave AI agents a year of a real software company's financial records and asked them to close the books month by month, as an accountant would, with each month's work carried into the next.

In the first month, the strongest agents produced books close to those prepared by qualified accountants. As the months went on, errors accumulated, and after several months the agents' overall balances differed from the accountants' by more than 15%. The benchmark reports that agents could defeat the reconciliation checks by inventing transactions or pulling in unrelated ones, and it shows one model searching for unrelated transactions that add up to the required amount.

The benchmark comes from a company and has not been peer reviewed.

Source: Penrose (2025).

A broader benchmark that simulated a small company points the same way. Across 175 tasks in software development, project management, data science, administration, human resources and finance, the best agent tested completed about 30% of tasks fully on its own (Xu et al., 2025).

Permissions decide how far such errors can spread. A system that proposes a query for an auditor to approve carries different consequences from one that runs the query, excludes records and acts on the result. Greater autonomy lets longer sequences run without interruption, but it also reduces the points at which a person sees the work in progress and can redirect it.

Reliability still trails peak AI capability

These systems are improving quickly, but what they can do reliably remains well behind what they can sometimes do. Researchers who measure progress by the length of the tasks that AI systems can complete report that this length has been growing steadily. However, the tasks that systems complete four times out of five are four to six times shorter than the tasks they complete half the time. Success rates are also lower on messier tasks, such as those that are novel or set in changing conditions (Kwa et al., 2026).

For audit, the stricter measure is the relevant one, and in this guide's judgment audit work needs reliability well above four times in five. If these trends continue, connected and agentic systems will become practical for more audit tasks. Whether they become dependable on audit work itself is a separate question, because much audit work is likely to be messier and more consequential than the tasks these benchmarks use. Their results therefore cannot show how soon agents will be reliable enough for it, and people are likely to remain essential at the points where judgment is exercised.

Contribution, risks and what leaders can do

CONTRIBUTION AND RISKS	When connected workflows work, they multiply what a team can do. When a step fails, however, the failure can pass unnoticed through everything that follows. Connected workflows therefore carry the greatest risk of the five kinds of work. The risk is highest where a system acts with wide permissions, on records that matter and in ways that make errors hard to detect and correct.
WHAT LEADERS CAN DO	Leaders can prepare now, without waiting for fully autonomous systems. First, they can connect tasks where each handoff can be checked, in workflows where people approve key steps, such as the scope of a search or the logic of a query, before the system proceeds. Second, they can design every connected workflow so that its intermediate results are visible and recorded, which allows auditors to see what happened at each stage as well as the final output. Third, they can grant permissions gradually, letting a system propose before it acts and expanding what it may do only as evidence of its reliability on the specific task accumulates. Institutions that build these habits on today's simpler workflows will be better placed to use more capable systems well as they arrive.

2.6 Lessons for leaders across the five kinds of work

Across the five kinds of work examined in this section, a consistent picture emerges of what AI offers audit, where it can mislead and what leaders can do about both.

AI's most promising contribution is reach. In each kind of work, AI can let an audit team do more than its resources would otherwise allow. It can help a team examine more of the evidence, analyze whole programs instead of a few files or a sample, consider more explanations before settling on one, produce clear reports more quickly and link these steps together. The evidence for these gains is strongest for defined tasks and specific tools, and on some tasks AI assistance produced less accurate results. Much of the evidence comes from professional work close to audit, and less of it from audit itself.

The weaknesses that matter tend not to show in the output. The output itself does not reveal what a search left out, which cases a model failed to flag, whether an explanation's reasoning holds, whether a summary dropped a qualification or whether a step in a workflow failed without notice. Fluent writing, accurate citations and good average performance can coexist with all of these. For this reason, AI's outputs cannot be judged by how convincing they look, and the auditors who check them need to understand the evidence well enough to notice what is absent.

Whether a weakness matters depends on what the output will be used for. A tool that suggests material for auditors to review can be valuable even when it is often wrong, because people screen what it produces. In contrast, a tool whose output decides what gets examined, appears in a report or acts on records needs much stronger evidence of its performance, because its errors reach further. The same tool can therefore be appropriate in one role and inappropriate in another. For leaders, the practical principle is to decide what role an output will play before deciding how far to rely on it.

Six working principles follow for audit leaders and engagement teams considering any of these uses.

- **Start with an audit need, not a tool.** Identify where gathering, comparing, explaining or reporting currently limits the work, and choose assistance that addresses that limit.
- **Begin where checking is easiest.** Uses whose outputs auditors already screen, or whose logic can be explained, build experience and confidence at low risk.
- **Match the evidence to the role.** The closer an output comes to defining the evidence, stating a finding or taking action, the more the institution needs to know about how it performs on that specific task.
- **Look for what is missing, not only for what is wrong.** Build checks for omissions into every use, such as testing whether known relevant material is found or examining cases a model did not select.
- **Keep the work visible.** Design uses so that sources, intermediate steps and the basis for each conclusion can be inspected.
- **Expect change and reassess.** The tools are advancing quickly, so judgments about a tool should be revisited as it changes, always against the audit task it serves.

Section 3 shows these working principles in action within a single audit. Sections 4 and 5 then examine how an institution supports AI-assisted conclusions and develops its auditors' judgment, before Section 6 synthesizes the wider shift, sequences the guide's six recommended actions across the first year and looks ahead at how public-sector audit could evolve.

3. Following one question through a performance audit: an illustrative example

Section 2 examined five kinds of audit work one at a time, and most of its evidence concerns single tasks. In a real audit, those tasks are joined, and what the team learns at one step changes what it asks and checks at the next. Following one question through a complete performance audit of a national affordable-housing fund shows how those steps connect and where AI adds value along the way.

The audit is fictional. The program, figures and outputs were constructed for this guide and supply no evidence about how AI performs, which is the subject of Section 2. This section is a shortened account of a wider four-question audit set out in full in Annex A, covering completion against schedule, alignment with housing needs, household access and continuing affordability. To keep the narrative focused, this section follows only the first question and refers to Annex A for the remaining questions, example tool classes and full assessment tables.

3.1 The audit and its question

The fund supports housing providers through agreements that set completion dates and release payments at milestones. Its progress report states that 8,400 homes have been delivered, without defining "delivered". The audit asks how many homes due by the end of the review period have actually been completed and what explains any delays. At the outset, department officials attribute the delays mainly to the limited capacity of housing providers.

The audit institution authorizes AI assistance for this engagement under conditions it sets in advance, including approved security arrangements for confidential program and provider records. An analyst works alongside the audit team to configure extraction and matching, and the team records what each use contributes and what checking it requires. Responsibility for the audit's conclusions stays with the engagement director.

The audit moves through five steps, and each one changes the next. Exhibit 3.1 gives an overview of the entire arc, and reading down its last column shows how the audit's question shifts from planning to reporting. Sections 3.2 to 3.6 then tell the story of each step, closing each one with what auditors checked and what that step shows.

Exhibit 3.1. Each step of the audit changes the next

Step	What the team does	What AI adds	What auditors check	What changes next
Planning	Assembles research and early program records	Finds and extracts sources, and drafts a planning brief	The brief, which treats the 8,400 homes delivered as homes completed	The team decides to establish what "delivered" means in every agreement
Preparing the records	Compares all agreements, provider reports and permit records	Extracts definitions and dates, and matches records that describe the same project	A sample of extracted values, all uncertain matches and a record missed by retrieval	The question narrows to Region C, where 800 of the 2,000 homes due are complete
Fieldwork	Interviews providers in Region C	Transcribes interviews and groups passages under categories the auditors define	The groupings, including categories that split one issue	The team obtains municipal application records through project files and voluntary cooperation
Analysis	Builds a timeline for each delayed project	Extracts dates and proposes a comparison the team had not made	Whether the timelines and the proposed comparison support each explanation	The explanation shifts from provider capacity to approval times that the fund's schedules did not allow for
Reporting	Drafts the report from the agreed findings	Drafts text and compares the drafts with the verified figures	One summary that states a wrong figure for Region C	The report states what the fund's progress report could not

Note: Constructed for this guide.

3.2 Planning: an error in the brief raises the first question

The audit team begins by gathering prior research on affordable-housing delivery alongside the fund's progress reports, a sample of funding agreements and municipal permit records. AI search tools work alongside conventional searches to locate studies, earlier audits and guidance. Extraction tools then place each study's setting, method and main finding into an evidence database, keeping every passage linked to its source (Section 2.1). From this assembled material, a generative AI assistant drafts a planning brief that proposes audit questions, candidate cases and early lines of inquiry.

The AI-drafted brief reads smoothly, draws on a wide range of material and proposes useful follow-up questions. However, it opens by stating that the fund has delivered 8,400 homes and that most regions are at or ahead of schedule. When the audit manager compares it with a brief written by an experienced auditor from the same package, a sharp discrepancy appears. The auditor's brief points out that occupancy permits record only about 5,700 completed homes.

Every single figure in the AI brief appeared accurately in the record it cited. The mistake lay in what the brief inferred from those figures, since it treated "delivered" as meaning "completed". The team keeps the AI brief's useful inquiries, corrects its account of delivery and concludes that a sample of agreements cannot settle what

"delivered" means across the program. It decides to compare every funding agreement.

What auditors check	The manager tests the AI planning brief side by side against one written by an experienced auditor from the same package and catches the definitional error behind the 8,400 figure.
What this step shows	A draft can cite every source accurately and still draw the wrong conclusion from them. Review therefore has to examine reasoning and definitions as well as the relevance, validity and reliability of the evidence (INTOSAI, 2019b, ¶106–108). When the manager scores both briefs against five criteria, the AI-drafted brief receives the higher total despite its fatal error (Exhibit 3.2). Annex A reproduces the full text of both briefs and sets out the test in detail (Annex A.2 and Exhibit A.3).

Exhibit 3.2. Why an aggregate score missed the error in the AI planning brief

Dimension, scored 0–2	Brief A (AI-drafted)	Brief B (auditor-written)
Sources cited beside statements	2	2
Breadth of material used	2	1
Interpretation essential to the delivery-planning decision	0	2
Specific follow-up	2	1
Clear organization	2	1
Total	8	7

Note: Constructed for this guide. Full text of Brief A and Brief B appears in Annex A (Exhibit A.3).

3.3 Preparing the records: the headline combines two different things

To test what "delivered" actually means across the fund, the team needs to compare every funding agreement with provider progress reports and municipal occupancy permits. Those records sit in different formats and use different project identifiers. Auditors define the exact fields they need, including the definition of delivery and the scheduled completion date, and an analyst configures an extraction and matching workflow. A language model extracts candidate values with their source passages, and matching methods link the records that belong to the same project (Section 2.2). That setup takes real front-loaded effort, which the institution records so it can weigh the cost at the end of the audit.

Once the links are checked, an analyst calculates the totals through explicit conventional rules. The full comparison confirms what the small planning sample could only suggest. One region uses "delivered" to mean completed homes with occupancy permits, while the other two regions use it for the milestone payment made when construction starts. The national headline of 8,400 delivered homes

therefore adds together two different achievements, only 5,700 of which are finished homes. The gap is widest in Region C, where only 800 of the 2,000 homes due have been completed. The audit's central question shifts from whether the fund is on schedule to why Region C is so far behind.

What auditors check	The team reconciles the agreement collection against the program's funding register, then checks a sample of extracted values, every uncertain project match and a sample of matches that look straightforward. While verifying Region C files manually, an auditor also notices two relevant agreement addenda that queries against the evidence database had missed. The team checks for similar missed addenda before finalizing the totals, adds the missing records by hand and logs the retrieval miss for review after the audit.
What this step shows	The calculation itself is ordinary analysis. What AI contributes is the preparation that makes full coverage practical within the audit's resources, revealing a definitional split that a sample could only hint at. At the same time, catching a missed addendum shows why retrieval from an evidence database still requires checks for omissions. Annex A gives the regional figures and explains how the records were prepared (Annex A.1, Exhibit A.1 and Annex A.2).

3.4 Fieldwork: providers point to the approval process

Using the linked program records, the team chooses its Region C fieldwork cases deliberately rather than from a handful of convenient files. It selects delayed projects alongside projects that finished on schedule so it can compare what differed between them, and interviews their housing providers.

With the interviewees' consent, a speech-recognition tool transcribes the conversations inside the approved environment, keeping every transcript linked to its recording and project file. A qualitative-analysis tool then groups interview passages under categories that the auditors define. Set side by side, the passages bring out a recurring pattern that providers describe in different words: municipal approval packages sent back, missing drawings and repeated requests for information.

Department officials had pointed mainly to weak provider capacity. The linked interviews point directly to the municipal approval process as well, giving the team a clear reason to examine project application histories. Because municipalities sit outside a federal audit office's direct statutory mandate, the team obtains those histories through the fund's project records and voluntary cooperation with the two largest municipalities.

What auditors check	Auditors read the proposed interview groupings against the original transcripts and correct categories where the tool split one approval issue across two labels.
What this step shows	Linking and grouping interview material helps auditors spot patterns across accounts that use different words for the same bottleneck. At this stage, however, providers' accounts are testimonial evidence of what they experienced, while establishing when delays occurred and what caused them requires corroboration from municipal application records. Annex A sets out how the interviews were selected, recorded and analyzed, and applies the same method to the audit's question on household access (Annex A.3).

3.5 Analysis: the explanation changes

The team builds a timeline for each delayed project and sets out what each explanation predicts. If provider capacity is the cause, applications should be late or incomplete. If municipal processing is the cause, there should be long waits after a complete application.

Extraction tools draw application, approval and construction dates from the records, and analytical software orders the events and calculates the intervals. Asked which comparison is missing, the AI assistant proposes one that the team had not made, namely approval times for similar projects outside the fund (Section 2.3).

When the team tests the data, the explanation changes. In the largest group of delayed projects examined, providers filed complete applications on time and approval then took a median of 14 months, against the six months assumed in the fund's schedules. Using published municipal permit data and voluntary disclosures from the two municipalities, the team runs the comparison the AI assistant suggested. Similar projects outside the fund waited a median of 13 months, and planning records show that the six-month assumption came from a region with faster approvals.

Provider capacity cannot explain the largest group of delays, though it still explains a smaller group of projects with incomplete applications. The finding is that the fund's schedules assumed six months for approvals that usually took more than twice as long in these municipalities. The recommendations can now address each cause separately (INTOSAI, 2019b, ¶127), with the findings limited to the projects examined.

What auditors check	Auditors verify the extracted dates against project files, test each explanation against what it predicts and confirm the external municipal benchmark data before accepting the revised finding.
What this step shows	AI widened the set of explanations the team tested by proposing a comparison it had not made. The conclusion, however, rests on timelines, municipal data and planning records that the auditors tested against competing explanations. Annex A gives the number of projects and homes behind each explanation, and the projects that fit neither (Annex A.4, "Developing the delivery explanation").

3.6 Reporting: a check against the verified figures

The team drafts the report from the agreed findings. A generative drafting tool prepares text from those findings, the supporting evidence and instructions about audience and terminology (Section 2.4). An AI-assisted check then compares every draft and summary against the verified figures.

That check catches a summary that gives the wrong number of completed homes in Region C, and reviewers correct it before the text leaves the team. With the numbers verified and every proposed wording checked against the underlying analysis, the report can state three things that the fund's progress report could not:

- permits record about 5,700 completed homes, against 8,400 reported as delivered;
- the difference arises because regions report different milestones under the same word; and
- in the Region C projects examined, the largest group of delayed projects (covering about 450 homes) waited a median of 14 months for approval after complete applications, against the six months that the fund's schedules assumed.

What auditors check	Reviewers verify the flagged summary discrepancy, check all numbers against the verified evidence and compare each proposed wording with the underlying analysis.
What this step shows	A wrong figure in fluent text is easy to read past. Comparing every draft and summary against verified figures is a straightforward check, and it is the kind of repetitive comparison that AI can assist. Annex A describes the reporting step in full (Annex A.5).

3.7 What the audit institution decides

Because the team recorded what each use contributed and what checking it required, the audit institution can decide about future use on evidence from its own work. It makes three decisions, which summarize a longer assessment of each use set out in Annex A (Annex A.6 and Exhibit A.4).

- **Continue.** Assisted extraction, matching, interview analysis and drafting continue as approved uses. They made a comparison across all agreements and projects practical, and their outputs could be checked against sources.
- **Change.** AI-drafted planning briefs are checked against the underlying records for errors of interpretation before anyone relies on them, because confirming the citations in this audit's brief would not have found its error.
- **Test first.** During the audit, database queries missed relevant documents that auditors found by other routes, including two Region C agreement addenda uncovered during verification (Section 3.3). Before any finding relies on retrieval answers, the institution will test retrieval with questions whose relevant records are already known. Until then, those answers serve as leads for auditors to examine.

3.8 What the example shows about AI's contribution

Across the five steps, AI contributes to the audit in four ways, each of which Section 2 examined separately.

- **Coverage.** Assisted extraction and matching let the team compare every agreement, provider report and permit record, where a review by hand would have covered a sample. That full-program comparison established across every agreement which regions defined "delivered" as construction starts rather than completions, confirming the regional shortfall behind the 8,400 versus 5,700 discrepancy.
- **Connection.** Research, program records, interviews and municipal records were held together and linked to their sources. The team could therefore move from a headline figure to the agreements behind it, and from providers' accounts to the application histories that tested them.
- **Alternatives.** The AI assistant proposed a comparison that the team had not made. The team made it, and the result showed that long approvals were usual in those municipalities and that the fund's schedules had not allowed for them.
- **Consistency.** Drafting assistance, together with a comparison of the drafts against the verified figures, helped the report say what the analysis supported and caught a wrong figure before it was published.

These contributions built on one another. The comparison of every agreement made a deliberate choice of cases possible, the interviews gave the team a reason to obtain the municipal records and the records changed the explanation. The main gain therefore lay less in time saved at any one step than in the result, since the team asked a better question and reached a better-supported answer than the one officials first offered.

At each step, the contribution depended on work that remained with the auditors. They defined the fields and categories, checked extracted values and matches, corrected the planning brief and tested each explanation against the records. In this

respect, the example illustrates both conditions in this guide's argument. The team could support its conclusions because every figure and statement remained linked to its source and was checked in proportion to what depended on it. The auditors could challenge what AI produced because they understood the program and the evidence well enough to see that "delivered" did not mean "completed" and that the first explanation of the delays was incomplete.

The example is constructed, so it shows what an institution could examine and supplies no measurement of what AI adds. The wider point, that AI contributes most when the team joins the steps of an audit, therefore rests on this illustration and on the reasoning in Section 1.1. The studies reviewed in Section 2 examine single tasks, and none follows AI assistance through a whole audit. Section 4 examines how an audit institution makes and sustains such choices across teams and over time, while Section 5 examines how auditors develop the judgment that this work required at each step.

4. Managing AI use across the institution

Sections 2 and 3 examined what AI can contribute to audit work. This section turns to what an audit institution needs in order to rely on that contribution and sustain it over time. Managing AI depends on connected choices across the institution. Leaders must decide where to apply AI, how to verify and document its results and how to sustain and evaluate each application. They must also manage external support without weakening independence or confidentiality, decide when to change or stop a tool and determine what to disclose to audited bodies and legislatures.

The first two subsections concern a single application: choosing where to use AI and deciding how to verify and document its results (Sections 4.1 and 4.2). The remaining five concern the institution as a whole over time (Sections 4.3 to 4.7). Together, these seven subsections address the first condition in this guide's argument, supporting the conclusions, and underpin the first five recommended actions sequenced in Section 6.

4.1 The most promising applications lie where audits currently stop short

Every audit institution decides which questions it can pursue and how far. AI could loosen some of the limits on those decisions, and those limits are also the best guide to where AI should be applied.

Starting from audit needs

With the time and people available, an audit team may set aside a line of inquiry, test a sample where it would prefer to test every item, or draw a narrower conclusion than the subject deserves. AI's value lies in extending what auditors can examine and in freeing time for the work that depends on their judgment, such as gathering evidence first-hand (GAO, 2026b).

Section 2 concluded that institutions should start from an audit need and not from a tool. For the institution as a whole, a practical way to do so is to identify where recent audits stopped short. Lines of inquiry dropped for lack of time or staff, samples tested where full testing was preferred, and routine work that displaced harder questions all point to problems that matter for the institution's audits. Some audit institutions already work this way. Their auditors identify the operational problems and help define what a tool needs to do, and the institutions report that the resulting tools fit real needs better and are adopted more readily by staff (OECD, 2026).

Institutions tend to begin where the work is well defined and time savings are easy to see, such as summarizing, drafting, translating or routine rule-based checks (Office of the Auditor General of Canada, 2026; Dolcet Redondo, 2026). Most applications reported across audit institutions so far aim to make existing work more efficient rather than to produce audit evidence (OECD, 2026). Such applications are a

sensible way to learn how the tools behave and what support they require. The applications that would most strengthen the audit may lie elsewhere.

Testing each candidate against what it would improve

Each candidate application can be assessed against four possible contributions: i) reaching evidence the team could not otherwise reach; ii) supporting a conclusion that rests on firmer ground; iii) freeing time for more important work; and iv) teaching the institution something it needs to know. An application that offers none of these is unlikely to justify the effort of introducing it, however capable the tool. Concentrating effort where AI adds most value to audits is also the stated aim of the European Court of Auditors (European Court of Auditors, 2026).

The assessment needs to be made on the work the institution actually intends to do, because the same tool can help with one task and hinder another (Section 2.1). AI should also be compared with other realistic ways of solving the same problem (Treasury Board of Canada Secretariat, 2026). Better-organized records, an existing analysis tool or a narrower audit question may be sufficient. Likewise, AI applied to a single step of the work may deliver most of the benefit without changing the whole process. In some cases, AI makes possible work the institution could not realistically have done before. The relevant comparison is then between the smaller inquiry the institution could have managed and the fuller one that AI allows.

Prioritizing among candidates

Where several applications look promising, three considerations need to be weighed together. The first is how much the application would improve the audit. The second is what could go wrong if the tool's results were mistaken or incomplete, and how much verification would be needed to catch such errors (Section 4.2). The third is readiness, namely whether the necessary data are available in usable form and whether people are in place to set the tool up, maintain it and support its users (Section 4.3). The United States Government Accountability Office screens proposed uses on similar grounds. It considers readiness and expected return as well as fit with its mission, and gives priority to uses that advance its core work and deliver measurable results (GAO, 2026b). An application of modest value for which the institution is ready may be a better first step than a more ambitious one that depends on data the institution does not yet have.

Checking that wider coverage reaches the right records

Examining more data helps only if the data include the records that matter for the audit question. A tool can examine only the records it is given, which creates two risks. The first, described in Section 2.2, is that a tool trained or built on past audit findings keeps directing auditors towards areas that were examined before (OECD, 2022; Financial Reporting Council, 2026). The second is that full testing can give auditors the impression that they already have the complete picture. In one experiment, auditors who had tested every transaction did less to follow up when external evidence later suggested possible fraud (Li, Brazel and Gold, 2026).

Before relying on a tool to widen an audit, institutions should therefore establish which records the tool examines and which it leaves out; whether it could lead

auditors to overlook areas or warning signs it was not designed to detect; and which parts of the audit auditors must still examine themselves.

Piloting on real audit work

Before committing to an application, institutions can pilot it on real audit work, at a scale that reflects how much would depend on it (GAO, 2026b). The pilot should test whether the application delivers the improvement for which it was chosen. Where the aim is to save time, the pilot should count all the effort involved, including preparing records, checking results, correcting errors and overseeing the tool, and not only the time the tool itself saves (European Court of Auditors, 2026). Where the aim is a stronger conclusion, the pilot should show what the team found that it would otherwise have missed. A pilot that reveals limits has also served its purpose. In a pilot within Barcelona's municipal audit function, for instance, the models tended to oversimplify when carrying out repetitive instructions, which showed the importance of defining test conditions and benchmarks carefully (Dolcet Redondo, 2026).

4.2 Reliance on AI results depends on the task, not only on the tool

Once an application has been chosen, institutions need to determine how far auditors can trust what the tool produces. Across an institution, this becomes a standing set of decisions about which tools may be used for which tasks, and how much verification each task requires.

Approving tools for specific tasks

Approval should cover a tool for a particular task, because a result can be accurate and still mislead if it is treated as showing more than it does. A tool approved to help auditors find contracts worth reading has not been approved to support an audit finding. Similarly, a tool that accurately summarizes the lease terms in a contract has not reviewed the whole contract, and an auditor who treats the summary as a full review may miss other obligations (Financial Reporting Council, 2026).

Approval also needs to be revisited. When an application is extended to new work, the relevant question is what has changed: the task, the records the tool works on or the people using it. The tool should then be checked again for whatever the change affects, without repeating every test. Even where nothing has been changed deliberately, performance should be checked periodically against known cases, consistent with ongoing monitoring expectations (Financial Reporting Council, 2026; GAO, 2026b). As one internal auditor who tested a model against known past fraud cases put it in an interview, monitoring is "not a set-and-forget thing" (Schwezoff, 2026).

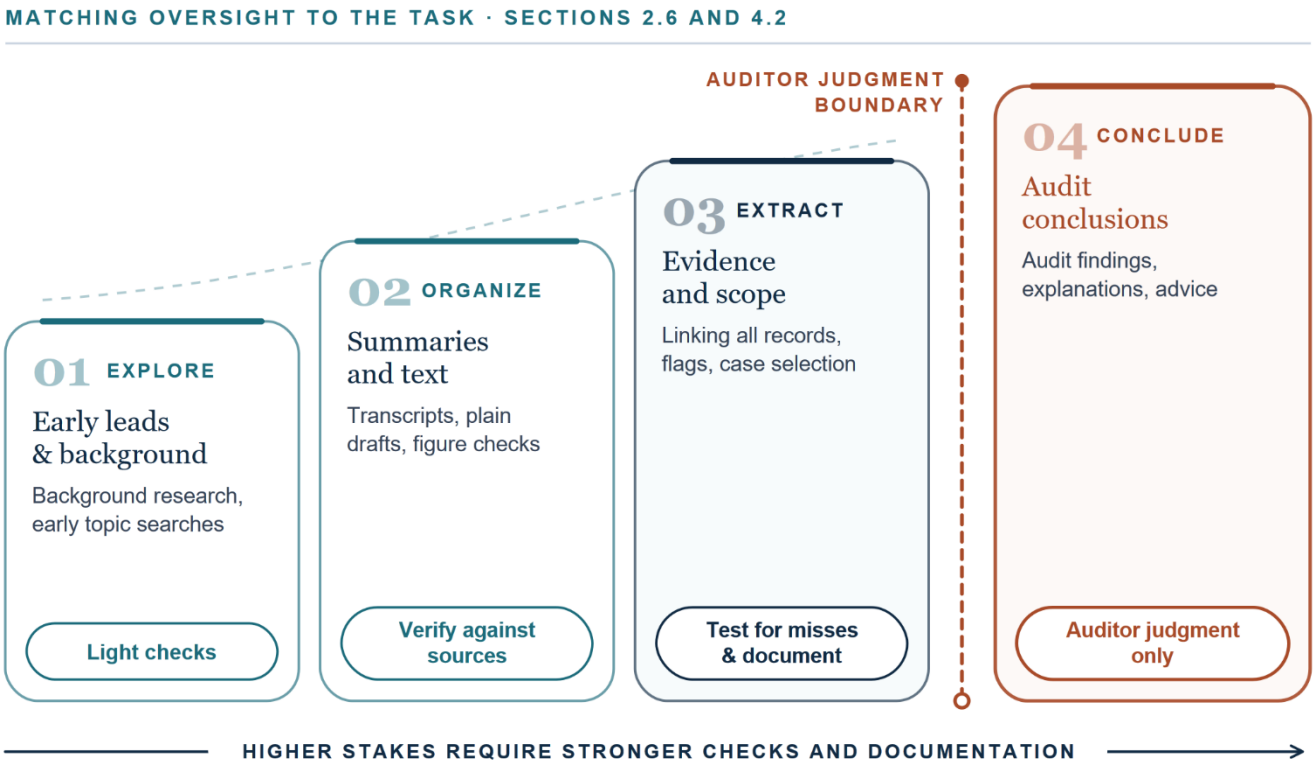
Matching verification to what depends on the result

Section 2 concluded that how much a weakness matters depends on what the output will be used for, and the amount of verification should follow the same principle. Verification should therefore concentrate on the errors that could change a conclusion. Two kinds matter most: i) evidence the tool missed and ii) results that have been read as showing more than they do.

Whatever the task, auditors must be able to confirm that results rest on reliable sources and are complete, and must be able to override them. Public audit institutions are building these expectations into their plans. These include piloting broad uses before committing to them, having people confirm that results are reliable, complete and free from bias, and ensuring staff can review and override a tool's results (GAO, 2026b; European Court of Auditors, 2026; Dolcet Redondo, 2026).

Figure 4.1 summarizes how oversight scales up as more depends on the result, up to the boundary where audit conclusions remain solely human judgment.

Figure 4.1. Matching verification and control to task stakes up to the auditor judgment boundary



Note: Approval covers a tool for a specific task, not in general. Responsibility for conclusions stays solely with the audit institution (CSAE 3001; ISSAI 100, ¶39).

Exhibit 4.1 then brings together, for each kind of result, the specific checks discussed in earlier sections.

Exhibit 4.1. Verification suited to each kind of AI-assisted result

Kind of result	Main check	What the check guards against	Discussed in
Search for sources	Compare with established search methods, or test whether sources already known to be relevant are found	Missing sources	Section 2.1
Answers drawn from a collection	Ask questions whose relevant records are already known, including contrary material, and examine the misses	A decisive record left out	Sections 2.1 and 3.7
Extracted values and matched records	Check a sample against the source passages, all uncertain matches and some matches that look straightforward	A misread figure or a wrong link	Sections 2.1 and 3.3
Calculation	Examine the source data, the rules and the code, and reproduce the result	Errors in data, rules or code	Section 1.2
Risk score or flag	Assess performance on the population being audited, and examine some cases the model did not select	Missed cases and bias inherited from past selections	Sections 1.2 and 2.2
Generated query	Test the query before relying on its results	A query that selects the wrong records	Section 2.2
Suggested explanation	Link each step to evidence, and state what evidence would show the explanation to be wrong	An inference that the evidence does not support	Section 2.3
Drafted or summarized text	Compare the final text with the agreed analysis	A lost condition, exception or degree of certainty	Section 2.4

These checks do not settle by themselves how far a result can be relied on. A test that finds the sources auditors already knew about, for example, says little about sources that nobody knew about. After the checks, the auditor therefore decides three things: i) which errors could change the finding; ii) what the tests showed about those errors; and iii) whether to rely on the result, test further or restrict the use.

Making review effective in practice

The way an application is set up can reduce the verification it needs, so that less falls on reviewers at the end of the process. A tool that has been well tested, and that trained staff use in a fixed way, may need less review than one that each user directs with their own instructions (Financial Reporting Council, 2026). Some applications also leave the auditor's own work in place. For example, a tool can summarize a board's minutes while auditors still read the minutes themselves and use the summary as a guide. In such cases, the auditor's own reading is the check. The verification an application requires is part of its cost and should be counted when deciding whether to adopt it.

The timing of review matters as much as its existence. Auditors exercise more judgment when they help select the evidence on which a conclusion rests than when they are asked to approve a finished answer. The developers of the New South Wales prototype described in Box 2.1, which let auditors select the evidence before the tool

drafted a response, warn that approving completed answers invites over-reliance (Commonwealth Scientific and Industrial Research Organisation, n.d.).

Moreover, a review step protects the audit only if reviewers look closely, and reviewers can become less careful when they believe a machine did the work. In one experiment, auditors checking identical work detected fewer planted errors when told a tool had prepared it than when told a colleague had. Prompting a questioning mindset beforehand reduced the effect (Peters, 2025). The study involved mostly early-career auditors on a simple task, so its result is best read as a warning. A review of 35 experimental studies of professionals working with AI describes the same tendency to favour automated recommendations over one's own judgment, and finds that experience can reduce it but does not remove it (Romeo and Conti, 2025).

A review of 106 experiments published between 2020 and mid-2023 found that, on average, a person working with AI performed worse than the better of the two working alone. The losses arose in tasks that involved making decisions, while tasks that involved creating content showed gains (Vaccaro et al., 2024). About 85% of the underlying experiments tested predefined-choice decision aids before widespread generative AI use, and results varied widely across tasks. The review therefore warns that adding a human reviewer does not automatically improve results, rather than measuring how current generative tools perform. Audit institutions report that training staff to remain sceptical of AI outputs, and to check them against original sources, has so far been their most effective safeguard (OECD, 2026). Institutions should also check what reviewers actually detect.

Documenting AI-assisted work in the audit file

Even a carefully verified AI-assisted result protects the audit only if the engagement file shows how the team reached its conclusion. Under Canadian and international public-sector audit standards, documentation must let an experienced auditor with no prior connection to the engagement understand the work (Auditing and Assurance Standards Board, 2015; INTOSAI, 2019a, ¶42; INTOSAI, 2019b, ¶86–88). The file must make clear what procedures the team ran, what evidence it gathered and how it reached its judgments.

Not every detail belongs on each engagement file. Guidance for financial auditors draws a practical line between institutional records and engagement working papers (Financial Reporting Council, 2025, pp. 5–6, 16). When an institution approves a routine tool for broad use, central records can hold its security review, baseline testing and version details. Where an engagement team uses AI in a bespoke way to select, extract, match or analyze evidence behind a finding, fuller documentation belongs on the engagement file itself.

For consequential uses, the engagement file should record five items:

- which tool, version and approved environment were used, linked to the institution's central approval record;

- what source records were provided to the tool and which records were excluded;
- the substantive instructions, queries or classification rules applied;
- what verification the team ran, what errors or missed records it found and how it resolved them; and
- the auditor's own reasoning linking the verified source records to the finding.

That documentation is not paperwork for its own sake. It enables an audit team to defend its evidence during clearance discussions with audited departments, internal practice inspection and questions before a legislative committee such as a Public Accounts Committee. Routine drafting or editorial assistance calls for much lighter records, because recording unnecessary detail does not by itself make an audit more reliable (Financial Reporting Council, 2025).

Review models are not yet settled

Audit standards already require auditors to review evidence and reasoning, and the same responsibilities apply when AI assists (Section 1.2). What is not yet settled is which combination of checks works for a particular AI application in public-sector audit. Each institution therefore has room to fit its arrangements to its own work and size. One requirement is constant: auditors must remain in control of the technology and be able to explain the key decisions behind their conclusions (International Forum of Independent Audit Regulators, 2026). Beyond that, practice among audit firms varies widely, and some firms use lighter approaches suited to smaller practices (International Auditing and Assurance Standards Board, 2026). Most of this guidance comes from financial audit, and public audit institutions will need to adapt it to their own mandates.

4.3 Sustaining AI requires continuing investment in people, data and records

An approved application remains sound only while the people, data and records behind it remain in place. Maintaining them is a continuing cost that needs to be planned from the outset.

Each application has three distinct requirements

Keeping an application sound requires three things: i) someone able to keep the tool running; ii) users able to interpret its results correctly; and iii) a record of why the institution decided to rely on it. Having one of these does not provide the others. A tool can be well maintained and still be misread by its users (Section 5 discusses how institutions build that judgment). Likewise, a record of an approval may explain the decision and still leave no one able to repair the tool. These requirements can separate even in an organization that uses AI well, as the audit of IP Australia shows (Box 4.1).

Box 4.1. IP Australia: capable use of AI with gaps in assurance

The Australian National Audit Office examined the use of AI at IP Australia, the agency responsible for patents and other intellectual property rights. It found that the agency used AI largely effectively and had strong controls on its main tool, although gaps in oversight remained elsewhere.

Several findings are relevant to audit institutions:

Restoring earlier versions. Earlier versions of a model were kept so that they could be restored. However, the agency told its auditors that restoring one was feasible but depended on manual steps and on the knowledge of individual staff. It had also not formally tested whether the model's results could be reproduced. The auditors suggested recording business approvals and selected changes as part of the agency's normal change-management process.

Demonstrating benefits. The quality of the work had been maintained and staff said the tools helped them work more effectively. Even so, the agency had not shown that the tools made it more efficient. Its estimates of benefits left out what the projects cost, and a benefits plan requested by its board had not been started.

Setting benchmarks. The tools had already moved beyond the trial stage by the time of the audit, and the auditors noted that benchmarks might have been easier to set during the trial.

An unused model. One model supplies examiners with patent applications when they have spare capacity. It had not been used since late 2021 because examiners had been fully occupied, while it cost about AUD 50 000 to maintain in 2024-25. The audit does not say that keeping it was wrong, since the tool was designed for use only when capacity allowed.

Source: Australian National Audit Office (2026), Artificial intelligence use in IP Australia.

Support is a continuing cost

The work of supporting an application continues for as long as the application does. Someone must maintain the tool, prepare its data and help the people who use it. Large audit institutions expect to need this kind of support. Their arrangements and plans include staff communities that share practice, stronger data management, training that is refreshed as the technology changes, basic training for all staff with deeper expertise for specialists, and updated templates and quality processes (GAO, 2026b; National Audit Office, 2026a). They also expect the work to become more demanding as small experiments grow into projects that need larger teams and sustained investment (European Court of Auditors, 2026). A study of audit institutions points in the same direction: across institutions with very different resources, the recurring obstacles are fragmented data systems, limited in-house technical expertise and the lack of shared measures of AI's impact (OECD, 2026).

Spreading knowledge beyond one or two people

No important tool should depend on one or two people. Where knowledge is concentrated, a resignation or reassignment can leave an institution with a tool it can neither maintain nor explain, as IP Australia's reliance on individual staff to restore earlier model versions illustrates (Box 4.1). How the tool works, how auditors use its results and why it was approved should therefore all be documented (OECD, 2026). Interviews with internal auditors point to the same risk. In one team of ten, a manager estimated that perhaps one and a half people really understood the team's

machine-learning work, and worried that the models might not be sustained if those people left (Schwezoff, 2026). Internal audit teams also often rely on hired or borrowed analytics staff, which makes their knowledge hard to transfer. These are concerns raised in interviews, and they point to a risk that can be addressed early.

Planning data work around priority audits

Data work varies with the type of audit and is often larger than it first appears. Some financial audits draw on datasets that remain stable from year to year. In contrast, performance audits often rely on documents from many organizations that arrive in no common format (European Court of Auditors, 2026). Much of an audit institution's knowledge may also sit in audit files rather than in structured data. One approach is to take stock of the institution's data in phases, starting with what the most valuable applications need. Another is to require each application to record where its data came from, how they were changed and which external systems were involved (GAO, 2026b).

Building on existing staff

Staff who look after data, audit methods and quality are natural starting points, provided they are given time to support others. Support also needs to reach everyone a tool affects. This involves identifying which groups of staff each tool affects, building a network of colleagues who can help others, sustaining training as the work changes, and clarifying how roles change and where problems should be raised (Treasury Board of Canada Secretariat, 2026). Audit institutions report two further lessons. Product owners who combine audit experience with digital skills help connect technical development to what auditors need. In addition, training fades unless staff apply it soon afterwards, which is why one institution requires participants to start a project using AI once their training ends (OECD, 2026).

For a national office, such arrangements can be modest. Brazil's Tribunal de Contas da União, for example, reports a dedicated AI unit of five staff and nine outsourced workers, working mainly on its internal assistant, along with regular internal events where staff share their experience (Tribunal de Contas da União, 2024). Smaller institutions can also gain real value from simpler analytics, controlled document workspaces and peer-to-peer sharing without building large specialist teams (International Auditing and Assurance Standards Board, 2026; OECD, 2026).

Matching governance and operating models to office size

Clear decision rights matter at every scale, so staff know who approves a tool, who checks security and who owns engagement-level judgment. Across most audit offices, those responsibilities divide into four roles, though in smaller offices one senior leader often carries more than one:

- **Auditor General and executive team:** set overall AI policy, risk tolerance, funding, external transparency posture and major procurement or partnership commitments.
- **Methodology and quality-assurance leads:** approve tools for specific audit tasks, define verification and audit-file documentation rules and lead periodic re-evaluation.

- **IT, security and privacy leads:** review data handling, tenancy isolation, vendor terms, model updates and ongoing technical support.
- **Engagement directors:** decide whether approved tools suit a specific engagement, oversee verification on the audit file and remain accountable for audit conclusions.

Public audit institutions vary widely in size. In Canada alone, legislative audit offices range from municipal and territorial offices with fewer than 50 staff to provincial offices and the federal Office of the Auditor General. Smaller offices do not need a large specialist unit to use AI responsibly. Exhibit 4.2 sets out proportionate operating models and recurring cost drivers across three office sizes as practical guidance.

Exhibit 4.2. Proportionate AI operating models by audit office size

Office profile	Typical starting posture	Governance and support roles	Main recurring cost drivers
Small office (under 50 staff, such as municipal, territorial or small provincial offices)	Sanctioned productivity or controlled document-workspace tools; simpler analytics before custom AI; peer-to-peer sharing with other audit offices	Named deputy or methodology lead with part-time IT and security support; short list of approved uses	Licences, staff training, verification time and occasional external specialist advice
Mid-sized office (50 to 250 staff, such as most provincial offices)	Approved zero-retention enterprise tenant, controlled document collections and selected extraction or analysis pilots tied to priority audits	Cross-functional steering group; methodology lead as product owner; dedicated data or analytics specialists	Enterprise tenant and security controls, data preparation, staff coaching and periodic retesting
Large office (over 250 staff, such as federal or national audit offices)	Portfolio of approved enterprise tools, targeted custom workflows (such as source-linked evidence databases) and structured multi-audit pilots	Formal AI governance council, dedicated digital and data unit, embedded methodology leads and annual executive reporting	Dedicated technical staff, cloud or local infrastructure, vendor assurance, change management and lifecycle monitoring

Note: Constructed for this guide as practical guidance for audit leaders, not empirical cost benchmarks.

Routing recurring problems to those who can resolve them

Staff who use a tool see its failures first, but they are not always best placed to identify the remedy. In one study of junior consultants who had just been given access to generative AI, the juniors mostly proposed managing its risks by changing how people worked. Expert recommendations published at the time also addressed the data, the model and the design of the system (Kellogg et al., 2025). A recurring problem should therefore reach someone who can change the tool or the audit method.

Overall, the evidence in this area consists mostly of plans and practitioner accounts, with little long-term evidence on what works best. Each institution will need to judge the appropriate scale of support for itself.

4.4 Institutions need to know which applications are in use and whether they deliver value

Once AI is in everyday use, institutions need a current view of which applications are in use, and a separate judgment on whether each one improves audits enough to justify the expenditure, staff time and verification it requires.

Maintaining a current view of AI use

Institutions should maintain a current view of which audits and decisions depend on AI, how staff are using it and who is responsible for each application. This view can become outdated quickly. Across 11 United States federal agencies, the number of reported AI uses nearly doubled in a single year, and the number of generative AI uses rose about nine-fold (GAO, 2025).

Approval alone does not provide this view, since it shows what was agreed but not what happened afterwards. Two audits by the Australian National Audit Office illustrate the gap. The Australian Taxation Office had no clear central view of its own AI use (Box 4.2), while IP Australia used AI largely effectively but still had gaps in oversight (Box 4.1).

Box 4.2. The Australian Taxation Office: governing AI without a central view or advance criteria

The Australian National Audit Office examined the governance of AI at the Australian Taxation Office. Its findings bear on three aspects of managing AI over time:

Visibility. The agency had no clear central view of its own AI use and no clearly defined arrangements for approving models. It had not monitored how well the models it had built in the past year were performing.

Value. Some models reported very high success rates among the cases they selected. However, the agency's ongoing reporting did not compare them with the approach they replaced, so the figures could not show whether the models performed better.

Withdrawal. Because performance criteria had not been set in advance, the agency's processes did not establish when a model should stay in use, be withdrawn or be brought back. The agency nonetheless withdrew two models in less than a year. It stated that, in general, models may be withdrawn because better ones become available, because their results become less useful or because they cost too much to keep running.

Source: Australian National Audit Office (2025), Governance of artificial intelligence at the Australian Taxation Office.

A complete view also includes tools that staff use without approval. Such use can be widespread and risky, as an audit of the Ontario government found (Box 4.3). Pressure to move quickly adds to the risk. In one vendor survey of large United States companies, 47% of senior executives said their organization had skipped its own approval process for urgent deployments (EY, 2026).

Box 4.3. Ontario: unapproved tools and unverified supplier assurances

The Office of the Auditor General of Ontario examined the use of AI in the provincial government. Three findings are relevant here and in Section 4.5:

Unapproved use. Staff visited about 400 AI websites in five months, and a security tool rated about 60% of them unsafe. The ministry responsible had not yet set roles, guidelines or timelines for tracking uses after they were approved.

Supplier testing. Officials had not addressed gaps in a supplier's test reports for an identity-verification system. The auditors recommended that the government validate test results provided by AI suppliers.

Data location. The evaluation of AI note-taking systems for health care relied solely on suppliers' confirmation that data stayed in Canada, without supporting evidence.

Source: Office of the Auditor General of Ontario (2026), Use of artificial intelligence in the Ontario government.

Unapproved use rarely stems from bad faith. In practice, it grows when auditors face tight deadlines, see clear value in everyday AI tools and confront slow or unclear approval paths. Blanket prohibitions without a workable alternative are hard to enforce and can push routine use out of sight. Providing at least one safe, clearly approved baseline tool, paired with plain rules on what records may never be uploaded, usually reduces the pull of unapproved public websites while giving leaders visibility into real staff demand.

An inventory is useful when it shows what each tool does, where auditors rely on it and who is responsible for it, and when people with the authority to act on it actually review it. The aim is to keep information about each application and the authority to act on it in the same place. Several audit institutions have set up councils or steering groups for this purpose. These bodies review proposals, oversee implementation and monitor compliance with legal and ethical requirements (OECD, 2026). The United States Government Accountability Office also plans yearly reports to leadership on what it is achieving and learning from AI (GAO, 2026b). Institutions will organize this work differently, depending on their size and structure.

Judging value separately from whether the tool works

Whether an application justifies the expenditure, staff time and verification it requires is a separate question from whether the tool still works. A tool can run reliably, and staff can value it, while the institution still cannot show what it contributes to audits. Both Australian audits illustrate this (Boxes 4.1 and 4.2). The Australian Taxation Office did not compare its models with the approach they replaced, and IP Australia had not shown that its tools made it more efficient.

Judging value should start from what the application was meant to improve. As discussed in Section 4.1, that may be a stronger conclusion, a wider inquiry or something the institution needed to learn, and not only time saved. The application should then be compared with a realistic alternative, counting all the staff effort it requires, including preparing records, checking results and correcting errors. Time

savings should also be interpreted with care, because whether faster work strengthens audits or reduces costs depends on what the institution does with the time released. An application that strengthens conclusions may justify its continuation even if it saves little time.

Deciding at the outset what evidence will be needed

Comparisons can become harder once an application is part of everyday work. By then, the institution may have stopped using the previous process and no longer has a current measure of how that process performed. In practice, this means setting out, when a pilot is approved, how it will be judged, so that the right information is collected while the pilot runs. It also means measuring how the existing process performs before the new tool is introduced (GAO, 2026b; Treasury Board of Canada Secretariat, 2026). IP Australia's auditors made the same point (Box 4.1).

At the same time, early measures have limits. An application often works differently once people are used to it, so early results can mislead. Keeping the old process running alongside the new one also has a cost. The aim is to retain enough evidence to make a fair decision later.

Matching evaluation effort to what depends on the application

Evaluation is not always straightforward. AI is often embedded in larger processes, which makes its own effect difficult to isolate. Some work would be impossible or far too slow without AI, so there may be no previous process against which to measure it (OECD, 2025). Audit institutions report the same difficulty. The link between AI use and tangible outcomes is often indirect, delayed or hard to isolate, and improvements in audit quality can be real but hard to capture with current measures. Where firm measures are lacking, structured feedback from auditors can help show what an application is contributing (OECD, 2026).

Not every decision requires a formal study. A pilot can be justified by what it teaches the institution, and lighter requirements suit pilots, with fuller tracking of benefits once an application becomes part of everyday work (Australian National Audit Office, 2026). However, an application on which audits will depend, or which commits resources for the long term, calls for firmer evidence.

4.5 External support is often unavoidable, but responsibility and key expertise stay with the institution

Few audit institutions can supply from their own staff all the technology, specialist skills and data work that AI requires. The relevant question is therefore on what terms an institution obtains external support, and what it must remain able to do for itself.

Much AI already arrives through existing suppliers

Much AI now reaches an institution inside software it already uses. Office productivity suites include AI tools by default, and suppliers are adding AI features to other mainstream products and analysis tools (National Audit Office, 2026b; European Court of Auditors, 2026). Nearly all the audit institutions consulted by the OECD rely on commercial platforms for productivity and collaboration, and some

run their AI work on commercial cloud services (OECD, 2026). An institution that builds nothing itself is therefore already relying on suppliers' AI.

Different kinds of external support meet different needs. Suppliers provide models, tools and the computing power to run them (National Audit Office, 2026b). Other audit institutions can share what they have learned and sometimes their solutions, which can help reduce costs and avoid duplicated work (European Court of Auditors, 2026). Universities and other external experts can contribute specialist knowledge (OECD, 2026). Consultants can carry the heaviest technical work at the start while the institution's own staff build their skills, although external knowledge can be hard to transfer to permanent staff (OECD, 2026; National Audit Office, 2026b). Shared arrangements can also lower the cost of getting started. Audit institutions at an early stage of adoption sometimes run initial trials on shared government infrastructure (OECD, 2026), and public bodies can procure jointly (Treasury Board of Canada Secretariat, 2026). Legislative audit offices must, however, weigh the independence risks of executive-run shared platforms before placing audit work on them, as examined below.

Choosing between buying, building and partnering

The choice between buying, building and partnering depends on what each option costs and delivers, and on how much control the institution retains over its tools, its data and its conclusions. The United Kingdom's National Audit Office offers a reasonable starting point: adopt trusted, secure tools where they meet the institution's needs, and build or partner only where necessary to protect audit quality, manage risk or obtain functions the market does not offer (National Audit Office, 2026a).

Because AI models are replaced so quickly, some audit institutions have external providers build tailored solutions while the institution keeps ownership of the data and the code. Even so, the OECD study links long-term success to institutions building dedicated teams of their own. Without them, an institution risks becoming too dependent on external providers or unable to adapt its tools as its needs change (OECD, 2026). A small institution can meet the same need by assigning the responsibility to named staff and sharing specialists with other institutions (Section 4.3).

Retaining the expertise to assess suppliers and their claims

Whatever support it obtains, an institution needs to retain enough expertise of its own to judge what it receives. It needs people who can specify what the institution needs, deal with suppliers effectively and see past exaggerated claims, because depending on external suppliers for scarce technical skills is itself a risk (National Audit Office, 2026b). A supplier's description of its own product is not sufficient. Suppliers usually do not share the data on which their models were trained. Assurance from an independent third party is valuable but does not replace testing: guidance for audit firms notes that a firm may still choose to test a tool itself for the use it intends (Financial Reporting Council, 2026). The Ontario audit reached the same conclusion about suppliers' test reports (Box 4.3).

Institutions also need protection against changes to the tools on which they rely. Suppliers update their models and components, and an update can change how a tool performs. An institution that runs a tool itself can control whether and how updates are applied. An institution that uses a managed service can seek advance notice of updates, retest the tool after a material change and keep the ability to restrict its use or leave. An update that materially changes a tool's performance is a reason to check the tool again, in line with Section 4.2 (Financial Reporting Council, 2026).

Retaining responsibility and independence

An audit institution can use tools and specialist support that others provide, but it cannot transfer responsibility for its audit work or conclusions. Canadian and international audit standards are explicit: the audit office retains full responsibility even when relying on external experts or outside parties (Auditing and Assurance Standards Board, 2015; INTOSAI, 2019a, ¶39; INTOSAI, 2019b, ¶65). It can work with external partners while retaining independence in its judgment (National Audit Office, 2026a).

An institution must also preserve independence in both fact and appearance (INTOSAI, 2019a, ¶13; INTOSAI, 2019b, ¶21–22). That principle sets a sharp boundary for legislative audit offices. Running an audit office's AI tools on shared infrastructure operated by an executive IT agency creates a clear independence risk. That risk is acute when the office audits that agency, or when central government administrators can access tenancy logs, prompts or usage records. Peer-to-peer sharing among independent legislative audit offices, such as exchanging tested policies, methods or specialist advice through national or international audit networks, lowers costs without compromising constitutional independence.

Institutions should also consider how sharing a commercial supplier with audited bodies affects perceived independence. In United Kingdom central government, for example, one commercial assistant is the main approved AI tool across many departments (National Audit Office, 2026b). An audit institution adopting the same tool relies on the same supplier as many of the bodies it audits, which calls for clear governance and separation of access.

Protecting audit information

Audit information needs particular protection whenever external providers are involved, because an external service can reduce the institution's control over where sensitive records from audited bodies are held and processed. In Canada, federal, provincial and territorial audit statutes pair broad access to government records with strict duties to comply with applicable security requirements and oaths of secrecy when examining departmental files (for example, *Auditor General Act*, s. 13(3)), reinforced by professional ethics standards (INTOSAI, 2019a).

Performance audits routinely handle personal information, commercial-in-confidence contracts, Cabinet confidences (such as records excluded from disclosure under section 69 of Canada's *Access to Information Act*) and departmental records classified as Protected B or higher. Under the federal *Directive on Security*

Management, Protected B designates sensitive information whose unauthorized disclosure could cause serious injury (Treasury Board of Canada Secretariat, 2019). Legislative audit offices are independent officers of Parliament or a legislature, not executive departments, yet they must maintain equivalent or stronger independent safeguards while meeting statutory security duties for departmental records.

Most audit institutions consulted by the OECD also expressed concern about data sovereignty, meaning which country's laws govern where data are stored and processed (OECD, 2026). Institutions should therefore know where and by whom their data will be processed (National Audit Office, 2026b), verify suppliers' assurances rather than taking them on trust (Box 4.3) and build standard data-protection clauses into contracts (GAO, 2026b).

In practice, deployment choices fall into three tiers with distinct risk profiles:

- **Public consumer AI services.** Unsuitable for confidential, personal or unpublished audit records, because prompts and uploaded files may be logged, retained or used to train models outside the office's control.
- **Enterprise zero-retention cloud tenants.** Dedicated institutional tenants with contractual data-residency controls, no training on audit inputs and verified security assurances; often practical for unclassified material and, where independently vetted and legally permitted, Protected B records.
- **Sovereign, private-cloud or local hosting.** Models hosted inside the audit office's own network or a strictly isolated sovereign environment; higher in technical cost, but necessary where statutory secrecy, Cabinet confidences or very high sensitivity bar external cloud processing.

In practice, an office's approved internal tools often lag behind newer public models, creating pressure to work around restrictions (Section 4.4). Offices can respond by drawing a bright line around genuinely public material, such as tabled legislation, published departmental plans, public evaluations or general methodology research, which may be examined in approved public tools under clear office rules, provided the prompt itself and the combination of records do not reveal an unpublished audit topic or line of inquiry. By contrast, casually anonymizing Protected B or unpublished audit records to use in a public consumer tool is rarely safe. Even when names are removed, specific dates, dollar amounts, contract clauses or the fact of an active audit can re-identify sensitive information and risk breaching statutory secrecy.

Agreeing terms that preserve future options

The terms of an arrangement also determine what an institution can do later. Canada's federal roadmap for scaling AI asks organizations to assess a supplier's capacity to scale and to work with other systems while an application is being piloted. At renewal, it asks them to review the risks of depending on the supplier, to plan a secure transition or exit, and to settle how data will be retained or disposed of (Treasury Board of Canada Secretariat, 2026). Similarly, United States federal policy, as summarized by the OECD, includes contract terms that protect against

dependence on a single supplier, secure the government's rights to its data, and require ongoing testing and periodic evaluation of the service's value (OECD, 2025). Considering these questions at the outset keeps the choices discussed in Section 4.6 open in practice.

The evidence on external support consists mostly of strategies, guidance and plans, much of it written for audit firms or for audited government bodies. None of the sources evaluates how well particular arrangements work over time.

4.6 Decisions to keep, extend, change or stop an application require evidence, authority and the ability to act

Institutions need to be able to review each application, respond in proportion to what they find and, in particular, decide how far AI may act autonomously.

Deciding in advance how applications will be reviewed

Keeping an application under review requires three things: i) evidence that something has changed; ii) someone with the authority to decide what to do about it; and iii) the practical ability to carry out that decision. These are different requirements, and having one does not supply the others.

The United States Government Accountability Office plans for its AI council to recommend whether each use should be expanded, modified or ended, based on pilots and ongoing monitoring, and for the council and its governance board to consider suspending or ending uses that miss thresholds for cost, schedule or performance. It notes that those thresholds are still to be developed (GAO, 2026b). Without such criteria, it is unclear when an application should continue or end, as the audit of the Australian Taxation Office showed (Box 4.2). Likewise, guidance for audit and risk committees in United Kingdom public bodies suggests asking of each pilot whether it is clear how decisions to stop, scale up or redesign will be made, based on evidence (National Audit Office, 2026b).

Responses from improving to retiring a tool

An application can continue as it is, be improved, be extended to new work, be restricted to fewer tasks, be paused, be replaced or be retired. The most useful response may be to narrow what a tool may do, or to redesign how auditors use it.

Organizations do make such changes. The Australian Taxation Office withdrew two models in less than a year (Box 4.2). In a vendor survey, most executives whose organizations had carried out formal AI reviews reported significantly changing at least a quarter of their AI systems, and some reported pausing or stopping them (EY, 2026). The survey does not show whether the reviews caused these changes. A review can equally support renewal or expansion.

The same applies to safeguards. Governments are advised to revise or remove safeguards that no longer serve their purpose, partly because unnecessary safeguards feed the risk aversion that holds back useful innovation, and to apply safeguards in proportion to the risk of each use (OECD, 2025).

Keeping the practical ability to change or stop a tool

A sound decision can still be difficult to implement. Changing or ending an application may depend on whether the institution has kept the records it needs, can recover its data and can continue the work by other means. These in turn can depend on choices made when the tool was procured and set up (Section 4.5). Reverting to an earlier version can also be difficult, as IP Australia's reliance on manual steps and on the knowledge of individual staff shows (Box 4.1).

Keeping a tool should also be a deliberate decision that is made and reviewed. IP Australia's unused model, which the agency continued to pay to maintain, shows why (Box 4.1).

Rechecking a tool when its use is extended

Extending an application calls for its own evidence. An application that works well for one kind of audit, one team or one set of records may not work as well for another. Guidance for audit firms treats use on a materially different task as a reason to certify a tool again (Financial Reporting Council, 2026), consistent with the principle in Section 4.2 that approval covers a tool for a particular task. Testing at larger scale demands time and resources, and complexity grows as a tool is connected to more systems (National Audit Office, 2026b). The recheck can be limited to whatever the extension changes (Section 4.2). An application can also remain limited indefinitely if that is where it is most useful.

Setting explicit limits on how far AI may act autonomously

How far AI may act on its own requires a separate decision. Section 2.5 described agentic systems, which choose their own steps towards an objective, and showed how an error can pass unnoticed from one step to the next. Greater autonomy can make a system more flexible, but it may also carry greater risk (Financial Reporting Council, 2026), and failures can do more harm because people have fewer opportunities to intervene (International AI Safety Report, 2026).

Audit institutions are proceeding cautiously (Section 2.5). Governance elsewhere has not always kept pace. In a vendor survey of large United States companies, about half of the executives whose organizations use agentic AI said their governance framework had not yet been updated to cover it, and about a quarter said their organization could not detect unauthorized AI agents operating internally (EY, 2026).

Institutions should therefore set, for each application, what the system may and may not do, when it needs a person's approval, and how it should respond when something goes wrong, including when it should escalate a problem and how it can stop safely (Financial Reporting Council, 2026). These limits should be revisited as tools improve, although the evidence does not yet show when more autonomy would be justified. Using AI to automate whole processes may also require a fundamental rethink of audit methods and standards (OECD, 2026).

Building checks that the AI cannot alter

Checks built into the system around the AI can help keep its work under control, particularly as it takes more steps on its own. Conventional software can check key steps of the AI's work against clear rules, in the same way every time. Guidance for

audit firms describes rules-based components that check whether an AI's outputs meet set requirements, review built in at the right points by people or by such components, and a requirement that the system keep a record of its current state, its goal and its key intermediate results (Financial Reporting Council, 2026). The same guidance notes a trade-off: a detailed fixed workflow usually lowers risk, but at the cost of flexibility, and can be too rigid for some uses (Financial Reporting Council, 2026). How tightly to fix the steps is therefore a choice to be made for each application.

The checks must also be protected from the AI they check. In the accounting benchmark described in Box 2.2, some agents defeated their reconciliation checks, for example by pulling in unrelated transactions (Penrose, 2025). A check is useful only if the AI cannot alter what the check examines.

The evidence on lifecycle review and autonomy is thin in one important respect. No source reviewed for this guide describes a complete, tested arrangement through which an audit institution has restricted or withdrawn an application, and the evidence on agentic AI in audit consists of guidance and early exploration. Institutions will need to set these limits with judgment and revisit them as evidence accumulates.

4.7 Disclosure, transparency and holding the institution to its own standard

Managing AI inside an audit office also raises an external accountability question. Leaders must decide what the institution owes to audited bodies, legislatures and the public when AI assists its work. They must also recognize how internal practice affects the office's credibility when it audits AI in government.

Being open about how AI is used

Audit institutions need to decide in advance what they tell audited bodies, legislatures and readers about their use of AI. The European Court of Auditors commits to communicating proactively with the bodies it audits about its use of AI for data analysis whenever necessary, in order to ensure transparency and foster trust (European Court of Auditors, 2026). The United States Government Accountability Office publishes an inventory of its AI uses (GAO, 2026a). For audit firms, the international forum of audit regulators states that communicating how technology is integrated into the audit, the extent of its use and its effect on decisions supports the oversight role of audit committees and confidence in audit quality (International Forum of Independent Audit Regulators, 2026).

These sources do not settle what a public audit report itself should disclose. An audit institution can nonetheless set clear rules in advance on three points:

- when it informs an audited department that approved AI tools are being applied to that department's records;
- what it publishes in an institutional inventory of AI uses; and

- what a report's scope and methodology section tells legislatures and readers when AI materially shaped the evidence behind a finding.

The institutional and engagement records described in Sections 4.2 and 4.3 give the office the factual basis to answer those questions plainly, including during parliamentary committee hearings.

Applying the institution's own standard to itself

Finally, an audit institution should hold its own use of AI to the standard it applies to others. As Section 1 notes, auditing government AI systems is outside the scope of this guide, yet the credibility link is direct. The United States Government Accountability Office commits to holding itself to the same high standards by which it holds audited agencies when it invests in, develops or deploys AI (GAO, 2026b). An audit office that can demonstrate clear task approvals, verified data safeguards, engagement-file traceability and documented value for its own AI tools stands on firmer ground when it asks audited departments to meet those same standards.

Section 5 turns from the institution to its people: how auditors build and keep the judgment they need to direct, interpret and challenge work that AI helps to produce.

5. Developing the professional capacity to use AI well

AI adoption raises a question about the kind of audit institution that leaders are building. New tools may expand what an institution can investigate and explain, and that contribution depends on people who can direct that work, question its results and take responsibility for its conclusions. This section examines what the evidence shows about developing those people while AI changes their work, and what it leaves unresolved. In doing so, it addresses the second condition in this guide's argument, sustaining judgment, and underpins the sixth recommended action sequenced in Section 6.

Three terms are used throughout. The *work* is what an auditor produces with AI assistance. *Competence* is what an individual auditor understands and can do, including without the tool where the role requires it. *Capability* is what the institution as a whole can sustain through its people, methods and arrangements. The central point of this section is that the three need to be assessed separately, because an improvement in the work does not by itself demonstrate competence or capability.

Most of the research used in this section was carried out outside audit, with programmers, customer-support staff and other professionals. It is used because it measures what people can do after working with AI, as distinct from what they produce with it, which the audit research reviewed here does not measure. Its findings are best read as well-founded questions for audit institutions to examine in their own work.

5.1 Good assisted work does not show what its author understands

A finished product and the understanding behind it can diverge

Audit institutions have always judged their staff partly through their work, since a well-reasoned working paper was reasonable evidence that its author understood the reasoning. With AI, that inference becomes less safe. An assistant can help find records, compare accounts and draft a clear analysis, and a supervisor may receive a better product as a result. However, the product alone no longer shows how much of its reasoning the auditor understands. This matters when a new record challenges the conclusion, when the next assignment concerns unfamiliar material or when the auditor later reviews someone else's work.

The clearest evidence that the two can diverge comes from a randomized experiment with programmers, in which participants who practised an unfamiliar task with AI scored lower on a quiz taken afterwards without it (Box 5.1). The experiment shows that people can complete a task with AI and understand less of it than people who completed it without. It concerns learning something new, and not the loss of existing expertise.

Box 5.1. Shen and Tamkin (2026): practising with AI and learning less

Design. Fifty-two participants, almost all with four or more years of programming experience, were asked to complete two tasks using a software library that was new to them. Half were randomly assigned to work with an AI assistant and half to work without AI, using the task materials provided. All participants were told to work as fast as they could and had at most 35 minutes. Everyone then took the same quiz without AI.

Results.

The AI-assisted group scored lower on the quiz. The quiz tested understanding of concepts, reading code and finding errors. The largest gap was in finding errors.

The study found no conclusive effect on speed. All 26 AI-assisted participants finished both tasks, while 4 of the 26 working without AI did not finish the second.

Among AI-assisted participants, those who asked conceptual questions or requested explanations tended to score higher than those who only asked the AI to generate code. Participants chose these approaches themselves, so their prior ability and motivation may explain the difference.

Limits. The experiment was small and short, and participants were asked to work quickly. It measured what people had learned immediately after practising with unfamiliar material. It did not measure what they retained later, whether established skills erode, or how people perform in professional work.

Source: Shen and Tamkin (2026), How AI Impacts Skill Formation.

When AI assistance supports skill building

The experiment in Box 5.1 does not stand alone, and other results are more encouraging. In a study of novice learners in introductory programming training, access to an AI code generator did not harm performance on tests taken a week later. Learners with stronger prior knowledge did better on those tests when they had had access (Kazemitabaar et al., 2023). Similarly, in a customer-support firm, workers who had used an AI system continued to handle cases faster during periods when the system was unavailable (Box 5.2).

Box 5.2. Brynjolfsson, Li and Raymond (2025): an AI tool built from the practices of a firm's strongest workers

Setting. A customer-support firm introduced an AI system that suggested responses during chats. The system had been built using successful conversations and the practices of the firm's strongest workers.

Results.

Less experienced and lower-skilled workers improved both the speed and the quality of their work.

The most experienced and highest-skilled workers saw small gains in speed and small declines in quality.

During outages of the system, workers who had been using it remained faster than they had been before it was introduced.

Limits. The evidence comes from one tool, one firm and one occupation. The outage analysis measured speed, not whether cases were resolved well. Workers were also gaining experience over the same period, which could explain some of the lasting improvement.

Source: Brynjolfsson, Li and Raymond (2025), Generative AI at Work.

The customer-support study suggests two things. First, an AI tool can pass on the working methods of an organization's best people to its newer staff, because each response that the system suggested showed a newer worker how an experienced colleague would have handled the case. Second, some of the improvement lasted when the tool was unavailable. The study measured speed and quality, and it does not show whether the less experienced workers acquired the judgment of the experienced ones. Field experiments with software developers likewise report larger productivity gains for less experienced developers, although the differences between groups are not statistically conclusive (Cui et al., 2025).

The longest study reviewed here concerns patent lawyers. In a three-month randomized experiment, 133 practising lawyers at eleven firms either received an AI drafting assistant or worked without one. Those with the assistant produced better work while using it. After three months, all participants completed a task without AI, and those who had used the assistant did better on average. However, the gain was concentrated among senior lawyers. Junior lawyers showed no average gain, and their scores spread out, with more good results and more poor ones (Autor et al., 2026). The authors suggest that foundational expertise may be needed before assisted practice builds lasting skill. The experiment was funded by Google, used a prototype tool and lost about a third of its participants before the final task, so its results are provisional.

"Experience" also means different things in these studies, including time in the job, recent performance and familiarity with one particular task. An experienced auditor may be new to a subject or a method. Institutions weighing who benefits from AI, and who is at risk, should therefore consider familiarity with the task at hand as well as seniority.

The implication: assess the work and the person separately

Taken together, these studies show that the quality of assisted work and the competence of the person can move in different directions, although they do not show how often this happens, for whom or for how long. For an audit institution, the implication is to decide what each AI initiative is meant to achieve, namely better work, more competent people or both, and then to look for evidence of each. Strong assisted output is valuable in its own right, and evidence of what its author could do on the next, unfamiliar assignment needs to be gathered separately.

5.2 AI could change how auditors learn, for better or worse

Auditors learn through explanation, practice, questions and feedback

Interviews in six Dutch audit firms describe how auditors learn on the job. A senior colleague explains what a task is for, and the junior auditor then does the task, asks questions when something is unclear and has the finished work reviewed and corrected. Each step relies on the others, since an explanation means little until the auditor has tried the work, and feedback is possible only once there is work to review. The interviews also show how easily this process breaks down: when supervisors are busy they give less guidance, and junior staff hold back their questions (Dierynck et al., 2026). The study did not examine AI.

AI could affect each of these steps, in either direction. If an assistant does part of a task, the auditor practises less of it and makes fewer mistakes of their own to learn from. If the assisted work looks better, the reviewer has less to correct and the auditor hears less about what they got wrong. Professionals interviewed about their own AI use described exactly this: their output improved, they received less corrective feedback and they felt they had missed a chance to learn. Others, in contrast, described using the time that AI saved for more demanding work from which they learned more (Oviedo-Trespalacios et al., 2026).

Whether handing a task to AI helps or harms learning therefore depends on what the task was teaching and on what the auditor does instead. The research reviewed here does not identify which audit tasks staff must do themselves in order to learn, or which can safely be handed over. The value of a task also changes as a person develops. Assembling records, for example, teaches a new auditor how those records are organized and where they are unreliable, but it teaches little to someone who has done it many times.

Before delegating a task to AI, institutions should therefore establish i) what the task currently teaches the people who do it, ii) what feedback they get from it and iii) where they will now get that experience. Neither keeping every manual step nor delegating every task is a sound default.

What is known about learning while using AI

Research on how people learn while they use AI is at an early stage, but two findings are relevant here. First, the way a person uses the tool appears to matter. In the programming experiment, participants who asked the AI to explain concepts scored higher afterwards than those who only asked it to generate code (Box 5.1). Similarly, in an experiment with 211 undergraduates, access to AI raised scores on unaided tests, and a smaller gain remained a week later. The gain held for students who used

AI as a tutor and faded for those who had it write for them (Contractor and Reyes, 2026). Second, a finished piece of work can hide what its author did not understand. Observation of novice programming students found that completed work could conceal confusion, and that students who struggled often believed they had performed better than they had (Prather et al., 2024).

The studies reviewed here do not show that any method of learning with AI works reliably, whether a way of prompting, a required amount of unaided work or a training course. Institutions therefore need direct evidence of what a person has learned, since signs of effort or of following a prescribed routine are weak evidence.

A proposal: build AI tools that require auditors to think

These findings suggest that how a tool is used affects what its user learns. Institutions have some control over this, because the same AI model can be set up in very different ways. One arrangement hands the auditor a finished answer to accept or reject, while another has the auditor do part of the reasoning before the AI does its part. Institutions decide which steps the tool performs, in what order and what it asks of the person using it. They can therefore choose arrangements that keep auditors thinking critically as they work, and this guide proposes that they do so deliberately. The proposal has not been tested in audit, and most of the studies behind it come from other settings.

Four design choices have some support.

Ask for the auditor's judgment before showing the AI's. In an online experiment, people accepted incorrect AI advice less often when they had to make their own decision first, or had to ask or wait for the AI's suggestion (Buçinca et al., 2021). A prototype built with the Audit Office of New South Wales applies a similar idea to audit work: auditors inspect the source passages the tool has retrieved, remove those that are weak or irrelevant and reorder the rest, before the tool drafts an answer (Commonwealth Scientific and Industrial Research Organisation [CSIRO], n.d.; see Box 2.1). An experiment with audit seniors points the same way, using conventional software and not generative AI. A decision aid that led auditors through the evidence in a structured dialogue, without recommending an answer, reduced two known judgment biases in a fraud-risk assessment (Lombardi et al., 2023).

Show the reasoning and the sources, not only the conclusion. As noted above, participants who asked the AI for explanations scored higher than those who asked only for finished code (Box 5.1). A tool can likewise be set up to present the steps and source passages behind a result, and to ask the auditor to confirm or correct a particular step. A model's own account of its reasoning may not describe what it actually did. The steps shown should therefore come from the system's record of the passages it retrieved and the queries it ran.

Ask the auditor to explain. Because finished work can hide what its author did not understand, a tool or the workflow around it can ask the auditor to state in their

own words why a conclusion follows and what evidence would change it. That statement can then be passed to the reviewer.

Show how experienced colleagues reasoned. Earlier audit files, methodological guidance and review comments record more than conclusions. They also record how experienced auditors reached those conclusions, including which explanation they accepted and which they rejected, what evidence persuaded them and what they decided the available figures could not show. A tool could retrieve that reasoning when it is relevant to the task at hand, with a link to the original file. The customer-support system in Box 5.2 worked in a comparable way, since it showed newer workers how the firm's strongest workers had handled similar cases.

These designs have costs, because they add effort and people may not welcome them. In the online experiment, participants gave their lowest ratings to the designs that most reduced their over-reliance on AI, and the designs helped people who enjoy effortful thinking more than they helped others. Moreover, the designs reduced over-reliance without removing it (Buçinca et al., 2021). Institutions should therefore expect some resistance, and they should test each design before relying on it. The test is what the staff who use it can later do: explain their reasoning, detect an error and repair a weak conclusion.

Some uses of AI do not need to be designed this way. Using a tool simply to get work done faster is reasonable, provided the auditor learns the underlying skill in some other part of their work. In addition, AI is only one way to fill a gap in learning. Guidance that is easier to follow, easier access to senior colleagues or faster feedback on work may fill it as well or better.

Reviewers must be able to challenge what the tool produces

The design of a tool is only part of the arrangement. An auditor's work is still reviewed by a more experienced colleague, and that review does two things. It protects the quality of the audit, and it is where much learning happens, because the reviewer can show why an inference was weak and how the evidence changes it.

For review to do both, the reviewer must be able to challenge in practice what the tool produces. A reviewer may have the formal right to overrule an AI system and still find it hard to do so. For example, the OECD describes a public service in which officials could override a system's negative recommendation only by contacting a special working group (OECD, 2025). Although the example comes from outside audit, it indicates what an institution should check:

- whether reviewers know the subject well enough to recognize an error;
- whether they can see the sources and reasoning behind a result;
- whether they have the time to examine them; and
- whether they are free in practice to reject the result and put forward a better-supported alternative.

Institutions can also test whether review catches AI errors in practice rather than assuming that it does (Section 4.2). Methodology or quality-assurance leads can run periodic spot-checks of AI-assisted working papers or short calibration exercises with known or seeded errors. Such checks show whether reviewers look past fluent text to verify definitions, omissions and source links.

Competence should be assessed under conditions that match the role

The remaining question is how an institution would know whether these arrangements are producing competent auditors. What it needs to know about a person depends on the responsibility that person will carry. Some things an individual must be able to do unaided, while others can properly be done with the help of a team, a specialist or a tool. Where a role requires an auditor to act without AI, the auditor needs an opportunity to show that they can. Where the work is collective, assessment should allow the support the role normally has, and knowing when to seek specialist help is itself part of competence.

In practice, audit offices do not need formal exams to separate the work from the person. During routine working-paper review, a manager can ask an auditor to explain orally why key criteria were chosen, how source passages were checked and which alternative explanations were ruled out. Offices can also assign developing auditors occasional short scoping memos or evidence notes to draft without AI assistance, keeping assessment lightweight and grounded in real audit work.

Voluntary guidance from the United States National Institute of Standards and Technology (NIST) makes a related suggestion, namely that tests of human proficiency should be kept distinct from tests of what the AI system can do (NIST, 2024). Neither a strong assisted report nor years of service shows, on its own, what a person can do.

5.3 Sustaining expertise over a career: leadership choices where evidence is thin

The questions above become larger when viewed across a career and across an institution, where little direct evidence is available. The points below are therefore reasoned implications for leaders, supported by evidence where possible.

Replacing the experience that routine tasks once provided

If AI takes over activities that once introduced new staff to audit reasoning, the institution needs another route to that experience. One possibility is to give people substantive investigative work earlier, with access to knowledge and appropriate supervision, which could be better than preserving routine tasks only because earlier generations learned through them. Professionals do report using time released by AI for more demanding work, although released time does not guarantee that outcome (Oviedo-Trespalacios et al., 2026).

Decisions about progression should rest on evidence of the responsibility a person can exercise, alongside the work they have produced with assistance. The institution needs to know whether a person can interpret evidence, recognize a material uncertainty and pursue a useful line of inquiry. It also needs people who can renew

its knowledge, by noticing when earlier guidance no longer fits a new question, investigating the difference and explaining it to others. The evidence reviewed here does not support forecasts of recruitment or staffing, and it does not prescribe a career structure.

Accounting for senior reviewers' and coaches' time

The time of experienced staff is part of both the cost and the benefit of AI. Faster production may release their time for inquiry and coaching, or it may leave them with more material to correct. In a pilot reported by the OECD, for example, AI produced training-course content rapidly, but the content still required expert review, which is where bottlenecks in production already existed (OECD, 2025). Institutions should therefore consider how AI changes the workload of their most experienced staff as well as the output of their junior staff.

Two transition pressures deserve early attention. First, senior reviewers can quickly become overloaded if teams generate long AI-assisted summaries or drafts faster than reviewers can verify the underlying evidence. Second, audit offices often face uneven adoption, with some staff eager to experiment while cautious colleagues resist tools they do not yet trust. Setting clear team norms on what checking must happen before a draft reaches review, and pairing peer coaching with realistic timelines, helps manage both pressures without forcing uniform adoption.

Using the competence arrangements institutions already have

Audit institutions already have arrangements for these responsibilities. ISSAI 150 is the standard on auditor competence issued by the International Organization of Supreme Audit Institutions (INTOSAI). It requires a supreme audit institution to determine the competencies its auditors need, to support them through its human-resource practices, to establish pathways for professional development and to assess competence periodically. It also describes development pathways as including practical experience, self-reflection and meaningful individual feedback (INTOSAI, 2022). AI gives leaders a reason to connect these existing responsibilities with decisions about how work is assigned and which tools are used.

The sources carry different weight. ISSAI 150 sets requirements for institutions that apply it, whereas the NIST guidance is voluntary and the OECD material is policy advice.

Learning from the institution's own experience

Because results from other settings may not transfer, an institution's own experience is its most relevant evidence. The OECD advises that experiments with AI need defined measures of success, evaluation and a way to capture lessons (OECD, 2025). For professional development, this means examining three things together: i) whether the audit work improved, ii) what staff came to understand and can now do and iii) how much support from experienced staff was required. Existing work reviews and professional-development reviews offer a starting point, without the need for an elaborate new assessment system.

Long-term effects on expertise remain unknown

No study reviewed for this guide follows auditors through extended AI-supported work and then assesses their independent judgment. The existing studies each cover part of the question, namely immediate learning, test scores after a week, speed during outages, unaided work after three months in another profession and workers' own perceptions. The OECD likewise notes that the long-term cognitive effects of AI use are not yet well understood (OECD, 2025).

Both outcomes remain possible. People who learn well with AI, and gains that persist without the tool, point to beneficial development, while less practice, less feedback and less independent problem solving point the other way. Audit institutions should therefore treat the long-term effect of AI on their expertise as a question to monitor in their own work, since adoption rates, staff satisfaction and short-term productivity cannot answer it.

6. Synthesis, sequencing and the future of public-sector audit

The Executive Summary sets out this guide's four findings and six recommended actions. This closing section steps back to synthesize the shift across Sections 1 to 5 and arrange those actions into a first-year sequence. It then looks ahead at how rapid, unpredictable AI advances could reshape public-sector audit over the next decade.

6.1 The shift beneath the tools

Read separately, the preceding sections deal with familiar audit concerns: assembling evidence, testing explanations, checking quality, managing vendors and developing staff. Taken together, they point to a deeper change in where public-sector audit creates its greatest value.

Shifting the bottleneck from assembly to judgment

For decades, the binding constraint in performance audit was gathering and organizing scattered records. Teams spent much of an engagement locating files, reconciling definitions and sampling a small fraction of a program because reading everything was impractical. AI begins to loosen that constraint. When tools can search across agencies, extract comparable records and link passages to their sources, auditors can spend more of their time on the work that matters most.

Upstream, the decisive skill becomes framing sharp audit questions and defining what evidence would settle them. Downstream, it becomes testing causal claims against wider records, spotting what an analysis missed and explaining conclusions clearly to audited departments and legislative committees. The offices that gain most from AI will not simply finish existing checklists faster. They will redirect freed capacity toward harder questions, fuller program comparisons and stronger scrutiny of public outcomes.

Balancing reach and credibility over time

Every section of this guide reflects a practical balance that audit leaders will revisit continuously as technology evolves. On one side sits genuine opportunity: government programs are growing more data-intensive, more interlinked and more likely to use automated systems themselves, and AI gives audit offices new ways to examine them at scale.

On the other side sits the foundation of public audit's authority: parliamentary and public confidence that an independent human institution owns and can defend every finding it reports. Keeping reach and credibility moving together means adopting useful tools deliberately, keeping evidence traceable and ensuring auditors continue to build judgment through real inquiry. Leadership here is not a one-time gatekeeping decision. It is the ongoing craft of widening what an audit office can investigate while strengthening the trust behind its work.

How architecture, governance and people reinforce each other

The experience reviewed across this guide also shows how technical design, governance and staff development strengthen one another when built together. Clean, source-linked evidence architecture makes checking results faster and clearer for reviewers. Proportionate governance gives audit teams safe, approved tools and clear boundaries, encouraging open experimentation rather than ad-hoc workarounds. Thoughtful workflow design keeps auditors actively engaged in questioning and testing results, building competence through daily work. Treating any one of these elements in isolation limits what the others can achieve; aligning them turns individual experiments into durable institutional capability.

6.2 Turning principles into a first-year sequence

An audit office does not need a multi-year transformation program to begin. The six actions at the front of this guide work well when sequenced in three practical stages, so basic safeguards are in place as teams move into more consequential audit work. Exhibit 6.1 arranges that sequence across four leadership roles: the executive team, methodology and quality-assurance leads, IT and security leads, and engagement directors (Section 4.3).

Exhibit 6.1. Sequencing leadership action across the first year and beyond

Stage	Priority moves	Primary lead	Signs that the stage is working
Stage 1: Set a clear foundation *(Months 1–3)*	Issue plain guidance separating public material from confidential and Protected B (or equivalent) records; provide a safe baseline tool while checking independence on shared IT; identify where recent audits stopped short; agree on pilot criteria, interim disclosure rules (Section 4.7) and clear task-approval owners	Auditor General and executive team, with IT, security and privacy leads	Staff have clear guidance and a safe baseline tool; informal use of unapproved public sites drops; two or three priority audit needs are ready to pilot
Stage 2: Run bounded pilots and learn *(Months 3–12)*	Pilot one or two task-specific uses (such as scoping search, controlled writing assistance or a small source-linked collection) on closed or bounded live files; test for omissions and check what reviewers detect; record baseline effort and audit-file documentation	Methodology and quality-assurance leads, with engagement directors	Documented pilot record showing what improved, what errors or missed records surfaced, how review worked and total staff effort required
Stage 3: Scale selectively and grow capability *(Year 2 and beyond)*	Expand, refine or retire applications against agreed criteria; extend into wider source-linked evidence work where value is clear; retest applications after major vendor updates; track unaided competence and adapt early-career learning; publish institution-wide disclosure principles	Executive team, methodology, IT/security leads and engagement directors	Active inventory with named owners; regular lifecycle decisions to grow, adjust or sunset applications; evidence of strong review quality and junior skill progression

Note: Constructed for this guide as a practical sequencing template aligned with the governance roles in Section 4.3 and office-size models in Exhibit 4.2.

The same sequence scales smoothly across office sizes. A small municipal or territorial office can complete Stage 1 with a concise policy note, a named deputy lead and peer sharing with sister offices, starting on public material if its IT platform is shared with audited bodies. It can then run Stage 2 as one modest pilot on a controlled document collection (Exhibit 4.2). A larger provincial or national office can run parallel pilots across teams and formalize vendor assurance, tenancy isolation and annual executive reporting. In every office, starting on small, reversible uses lets teams build strong habits before stakes rise.

6.3 Looking ahead: an unpredictable frontier and how public audit could change

Sections 1 to 5 stayed close to what current research, audit reports and early prototypes can support. This subsection steps deliberately beyond that evidence. It examines what is taking shape at the frontier of AI development today, why the next five to ten years remain genuinely unpredictable and how those shifts could open new possibilities for public-sector auditing.

Frontier trends and strategic unpredictability

Three technical developments are expanding what AI tools can do, alongside real uncertainty about how fast reliability will catch up.

First, frontier models are extending both the reasoning depth and the length of multi-step tasks they can carry out. Benchmark tracking shows steady growth in the duration of structured technical work that models can complete autonomously (Kwa et al., 2026). Tools are evolving from single prompts toward agentic systems that can plan steps, query documents, run analytical code and check intermediate results. At the same time, real-world agent trials show that reliability still trails raw capability on messy, multi-step tasks where subtle errors can pass unnoticed (Kwa et al., 2026; Penrose, 2025; Xu et al., 2025). Progress also remains uneven across adjacent tasks, making human checkpoints essential even as tools grow more capable (Dell'Acqua et al., 2026).

Second, capabilities diffuse rapidly across the market, often faster than public-sector procurement cycles. Advances that appear first in frontier commercial services frequently reach mainstream office suites, lower-cost enterprise tiers and open-weight models within a year or two. That pace creates both a caution and an opportunity. It rewards modular designs over heavy, multi-year lock-in to a single proprietary system. It also means capable locally hosted models are becoming more realistic, potentially opening secure, sovereign analysis of confidential and Protected B records through dedicated or peer-shared infrastructure sooner than many expect.

Third, the longer-term trajectory remains genuinely uncertain. Authoritative international assessments note that neither AI developers nor governments can reliably predict how quickly real-world dependability will match benchmark gains (International AI Safety Report, 2026). Technical plateaus, changing compute costs, supplier shifts or new legal frameworks could each alter the path ahead. For audit leaders, that unpredictability is not a reason to pause. It points toward investing in

durable, model-independent strengths: clean source-linked evidence, modular workflows where underlying tools can be swapped, proportionate verification and experienced auditors.

Five ways public-sector audit could evolve

If frontier tools keep advancing over the next five to ten years, even unevenly, five deeper shifts could meaningfully reshape how public audit institutions work.

From isolated audits to cumulative institutional memory. Most performance audits still assemble a fresh body of evidence for one engagement and archive it once the report is tabled. As source-linked evidence bases become easier to build and maintain, audit offices could carry forward public records, reusable syntheses and legally authorized material across multi-year programs, recurring grants and follow-up audits, while respecting statutory secrecy and engagement-file boundaries. Institutional memory could grow from what veteran auditors remember into a living, inspectable evidence base connecting past findings, departmental commitments and current delivery across audit cycles.

A more multidisciplinary team and redesigned early-career learning. Traditional audit offices have relied on a broad base of junior staff conducting routine extraction, document review and testing, building judgment gradually through that work. As AI takes over more mechanical preparation, audit teams may evolve toward smaller, more multidisciplinary units blending senior auditors, data specialists and domain experts. That shift can narrow traditional entry-level roles, yet it also creates an opportunity to bring newer auditors into substantive inquiry earlier. To realize it, offices will need deliberate pathways for newer staff to build scepticism, pattern recognition and field judgment alongside automated tools (Dierynck et al., 2026; National Audit Office, 2026b).

Keeping pace with an AI-driven state. Although this guide focuses on using AI inside the audit office, internal capability and external oversight naturally reinforce each other. Government departments are embedding automated decision systems, risk scoring and generative tools across benefits, tax, procurement and service delivery. An audit office that builds firsthand experience testing AI omissions, model drift and supplier claims on its own engagements will be far better equipped to audit those systems in government and more credible when advising audited bodies.

Clarifying the human mandate as tools grow more autonomous. Agentic systems are moving closer to querying live government datasets, proposing analytical tests and assembling draft working papers. As capability grows, the central question shifts from what software can execute to what a democracy wants delegated on behalf of Parliament or a legislature. Deeper automation may eventually invite updates to assurance methods and standards (OECD, 2026). Through any such evolution, one principle remains constitutional rather than technical: citizens and legislators look to an identifiable, independent officer and professional team to stand behind public audit judgment. Defining how powerful tools serve that mandate will be a central leadership task of the coming decade.

Shared public-audit infrastructure. Fast model turnover, sovereign-security needs and competition for specialist talent make it hard for small and mid-sized legislative offices to build and test everything alone. Over the next decade, independent audit offices stand to gain substantially from closer horizontal collaboration. Through peer networks such as the Canadian Council of Legislative Auditors, the Canadian Audit and Accountability Foundation and INTOSAI, offices can pool supplier assurance reviews, share open verification benchmarks, exchange source-linked workflow designs and co-invest in specialist skills while preserving independence.

6.4 Building the profession's own evidence base

The most important questions about AI in public-sector audit will not be answered by software demonstrations, laboratory benchmarks or studies of other industries. They will be answered inside real audit offices working under real statutory confidentiality, entity-clearance discussions and parliamentary scrutiny.

Every office that runs a thoughtful pilot, tests what reviewers detect, tracks how early-career auditors grow and shares honest findings on both successes and limits adds evidence the entire profession needs. Exchanging those practical lessons across independent offices will matter as much as adopting any single model.

The opportunity ahead is substantial. Audit institutions that pair traceable evidence with proportionate verification, keep investing in human judgment and guard their independence can examine public programs more thoroughly and explain them more clearly than before, while standing confidently behind every conclusion they report.

References

Access to Information Act, R.S.C., 1985, c. A-1. https://laws-lois.justice.gc.ca/eng/acts/a-1/

Auditing and Assurance Standards Board. (2015). *Direct engagements* (CSAE 3001). CPA Canada Handbook – Assurance.

Auditor General Act, R.S.C., 1985, c. A-17. https://laws-lois.justice.gc.ca/eng/acts/a-17/

Australian National Audit Office. (2025). *Governance of artificial intelligence at the Australian Taxation Office* (Auditor-General Report No. 26 2024–25).

Australian National Audit Office. (2026). *Artificial intelligence use in IP Australia* (Auditor-General Report No. 43 2025–26).

Autor, D., Rodchenko, T., Martin, J., Iscenko, Z., Strand, S., Pearl, D., & Ferere, M. (2026). *Does AI assistance enhance or erode expertise? Evidence from a three-month field experiment in patent drafting* (NBER Working Paper No. 35720). National Bureau of Economic Research. https://doi.org/10.3386/w35720

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2). https://doi.org/10.1093/qje/qjae044

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188. https://doi.org/10.1145/3449287

Commonwealth Scientific and Industrial Research Organisation. (n.d.). *Meaningful human oversight in AI-assisted public-sector auditing: Collaboration with the Audit Office of New South Wales*. https://research.csiro.au/isp/research/privacy_mlai/meaningful-human-oversight-in-ai-assisted-public-sector-auditing-collaboration-with-the-audit-office-of-new-south-wales/

Contractor, Z., & Reyes, G. (2026). *Experimental evidence on the learning impact of generative AI* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.08849

Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2025). *The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers* [Working paper; published in *Management Science*]. https://doi.org/10.1287/mnsc.2025.00535

Dell'Acqua, F., McFowland, E., III, Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*. Advance online publication. https://doi.org/10.1287/orsc.2025.21838

Dieffenbach, K. R. (2026, February 25). *Statement of Kenneth R. Dieffenbach, Executive Director, Pandemic Response Accountability Committee, before the U.S. Senate Committee on Small Business and Entrepreneurship concerning "From fraud to recovery: Restoring integrity in small business programs."* Pandemic Response Accountability Committee. https://www.pandemicoversight.gov/dieffenbach-february-2026-testimony

Dierynck, B., Marangoni, C., Peters, C. P. H., & Weijers, L. (2026). *Human capital development in auditing: A qualitative study of learning mechanisms, challenges, and differences across audit firms* [Working paper, FAR Project No. 2023B04].

Dolcet Redondo, A. (2026, February 27). *Artificial intelligence in public auditing as a tool for improving management*. KDZ Centre for Public Administration Research. https://www.kdz.eu/en/news/blog/artificial-intelligence-public-auditing-tool-improving-management

European Court of Auditors. (2026). *Artificial intelligence strategy 2026–2030*. https://doi.org/10.2865/0385430

EY. (2026). *EY US AI risk and governance survey: Closing the confidence gap*. https://www.ey.com/en_us/newsroom/2026/09/ey-survey-finds-that-autonomous-ai-implementation-outpaces-oversight-yielding-an-ai-governance-gap

Financial Reporting Council. (2025). *AI in audit: Illustrative example and documentation guidance*.

Financial Reporting Council. (2026). *Generative and agentic AI guidance: Risks, mitigations and illustrative examples*.

Fitzpatrick, T., Kelly, S., Carey, P., Walsh, D., & Nugent, R. (2025). *Assessing generative AI value in a public sector context: Evidence from a field experiment* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2502.09479

Government Digital Service. (2025, June). *Microsoft 365 Copilot experiment: Cross-government findings report*. GOV.UK. https://www.gov.uk/government/publications/microsoft-365-copilot-experiment-cross-government-findings-report

HM Treasury & Evaluation Task Force. (2026, May 15). *Magenta Book: Central government guidance on evaluation*. https://www.gov.uk/government/publications/the-magenta-book/magenta-book-central-government-guidance-on-evaluation-html

International AI Safety Report. (2026). *International AI safety report 2026*. https://arxiv.org/abs/2602.21012

International Auditing and Assurance Standards Board. (2026). *Technology quality management roundtables: Outcomes and next steps*. https://ifacweb.blob.core.windows.net/publicfiles/2026-02/IAASB-Technology-Quality-Management-Roundtables-Feedback.pdf

International Forum of Independent Audit Regulators. (2026). *Use of technology in audits: Innovation and audit quality*. https://www.fsa.go.jp/ifiar/2026TTFReport.pdf

International Organization of Supreme Audit Institutions. (2019a). *Fundamental principles of public-sector auditing* (ISSAI 100). https://www.issai.org/wp-content/uploads/2019/08/ISSAI-100.pdf

International Organization of Supreme Audit Institutions. (2019b). *Performance audit standard* (ISSAI 3000). INTOSAI Framework of Professional Pronouncements.

International Organization of Supreme Audit Institutions. (2019c). *The performance auditing process* (GUID 3920). https://www.issai.org/wp-content/uploads/2019/08/GUID-3920-The-Performance-Auditing-Process.pdf

International Organization of Supreme Audit Institutions. (2022). *Auditor competence* (ISSAI 150). https://www.issai.org/wp-content/uploads/2022/11/ISSAI-150-Auditor-competence.pdf

Kazemitabaar, M., Hou, X., Henley, A. Z., Ericson, B. J., Weintrop, D., & Grossman, T. (2023). *How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment* [Author manuscript]. https://doi.org/10.1145/3631802.3631806

Kellogg, K. C., Lifshitz, H., Randazzo, S., Mollick, E., Dell'Acqua, F., McFowland, E., III, Candelon, F., & Lakhani, K. R. (2025). Novice risk work: How juniors coaching seniors on emerging technologies such as generative AI can lead to learning failures. *Information and Organization, 35*, Article 100559. https://doi.org/10.1016/j.infoandorg.2025.100559

KPMG Canada. (2026, October 13). *Canadian AI procurement a priority as public sector pushes ahead with adoption*. https://kpmg.com/ca/en/media/2026/10/canadian-ai-procurement-a-public-sector-priority.html

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Painter, C., . . . Chan, L. (2026). *Measuring AI ability to complete long software tasks* (Version 4) [Preprint]. arXiv. https://arxiv.org/abs/2503.14499v4

Lau, O., & Golder, S. (2025). Comparison of Elicit AI and traditional literature searching in evidence syntheses using four case studies. *Cochrane Evidence Synthesis and Methods, 3*(6), Article e70050. https://doi.org/10.1002/cesm.70050

Li, X., Brazel, J. F., & Gold, A. (2026). *An unintended consequence of full population testing on auditors' professional skepticism* [Working paper, FAR Project No. 2021B01]. https://doi.org/10.2139/ssrn.4989842

Liu, S., Geng, M., & Hart, R. (2025). Exploring generative AI techniques in government: A case study. *IEEE Intelligent Systems*. Advance online publication. https://doi.org/10.1109/MIS.2025.3569429

Lombardi, D. R., Brown-Liburd, H. L., & Munoko, I. (2023). *Promoting systematic reasoning: The effects of interactive decision support systems on judgment biases in professional decision-making* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.4318689

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*. https://doi.org/10.1111/jels.12413

National Audit Office. (2026a). *The NAO's AI strategy: The future of artificial intelligence in our work*. https://www.nao.org.uk/corporate-information/the-naos-ai-strategy/

National Audit Office. (2026b). *Good practice guide for organisations using AI*. https://www.nao.org.uk/wp-content/uploads/2026/05/good-practice-guide-for-organisations-using-ai.pdf

National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). https://doi.org/10.6028/NIST.AI.600-1

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Office of the Auditor General of Canada. (2026). *2026–27 departmental plan*. https://www.canada.ca/en/auditor-general/transparency/departmental-plans/2026-27-full-report.html

Office of the Auditor General of Ontario. (2026). *Use of artificial intelligence in the Ontario government*. https://www.auditor.on.ca/en/content/specialreports/specialreports/en26/2026_AI_EN.pdf

Organisation for Economic Co-operation and Development. (2022). *Modernising integrity risk assessments in Brazil: Towards a behavioural-sensitive and data-driven approach*. OECD Publishing. https://doi.org/10.1787/ad3804f0-en

Organisation for Economic Co-operation and Development. (2025). *Governing with artificial intelligence: The state of play and way forward in core government functions*. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/governing-with-artificial-intelligence_398fa287/795de142-en.pdf

Organisation for Economic Co-operation and Development. (2026). *The state of artificial intelligence in public audit: Evidence from selected countries and the European Union* (OECD Artificial Intelligence Papers No. 58). OECD Publishing. https://doi.org/10.1787/f4a6c658-en

Oviedo-Trespalacios, O., Laksanadjaja, F., & Torkamaan, H. (2026). Skill development, maintenance, erosion, and revaluation: How knowledge workers experience generative AI. *Global Issues Challenge: Challenges in AI at the Human Level, 234*, 55–64. https://doi.org/10.54941/ahfe1007458

Penrose. (2025). *Can LLMs do accounting?* [AccountingBench benchmark]. https://accounting.penrose.com/

Peters, C. (2025). *Auditor automation usage and professional skepticism* [Working paper, June 2025 version]. https://doi.org/10.2139/ssrn.4309348

Prather, J., Reeves, B. N., Leinonen, J., MacNeil, S., Randrianasolo, A. S., Becker, B. A., Kimmel, B., Wright, J., & Briggs, B. (2024). *The widening gap: The benefits and harms of generative AI for novice programmers* [Author manuscript]. https://doi.org/10.1145/3632620.3671116

PwC US. (2026, October 8). *The new internal audit—Moving beyond traditional constraints*. https://www.pwc.com/us/en/services/consulting/risk-regulatory/enterprise-risk-management-controls/internal-audit-services/ai-internal-audit.html

Romeo, G., & Conti, D. (2025). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI & Society*. https://doi.org/10.1007/s00146-025-02422-7

Schwezoff, D. (2026). *The current role of artificial intelligence (AI) in the changing internal audit methodology* [Master's thesis, Oulu University of Applied Sciences].

Shen, J. H., & Tamkin, A. (2026). *How AI impacts skill formation* [Manuscript, 3 February 2026]. https://arxiv.org/pdf/2601.20245v2

Si, C., Yang, D., & Hashimoto, T. (2024). *Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.04109

Treasury Board of Canada Secretariat. (2019). *Directive on security management*. Government of Canada. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32611

Treasury Board of Canada Secretariat. (2026). *Roadmap to scale AI projects*. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/g7-ai-challenge/roadmap-scale-ai-projects-public-sector.html

Tribunal de Contas da União. (2024). *Uso de inteligência artificial aprimora processos internos no Tribunal de Contas da União*. https://portal.tcu.gov.br/imprensa/noticias/uso-de-inteligencia-artificial-aprimora-processos-internos-no-tribunal-de-contas-da-uniao

U.S. Government Accountability Office. (2025). *Artificial intelligence: Generative AI use and management at federal agencies* (GAO-25-107653). https://www.gao.gov/products/gao-25-107653

U.S. Government Accountability Office. (2026a). *Artificial intelligence at GAO* [Use-case inventory as of March 2026]. https://www.gao.gov/science-technology/artificial-intelligence-at-gao

U.S. Government Accountability Office. (2026b). *Artificial intelligence strategy*. https://www.gao.gov/science-technology/artificial-intelligence-at-gao

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Wray, S. (2026, March). *Accelerating government transformation: Takeaways from AccelerateGOV 2025*. Global Government Forum. https://www.globalgovernmentforum.com/download/77241/?tmstv=1772555287

Xu, F. F., Song, Y., Li, B., . . . Neubig, G. (2025). *TheAgentCompany: Benchmarking LLM agents on consequential real world tasks* (Version 3) [Preprint]. arXiv. https://arxiv.org/abs/2412.14161v3

Annex A. The full worked example: using AI across a performance audit

This annex serves as a practitioner's technical companion to Section 3. Section 3 follows one of the illustrative audit's four questions through five steps so audit leaders can see how linked tasks change an inquiry. This annex sets out the full engagement for engagement directors, methodology leads, auditors and analysts, covering all four audit questions, example tool classes, data-handling safeguards, the full text of the planning-brief test and the detailed post-audit assessment.

The audit is fictional. The program, the figures and the outputs are constructed for this guide and add no evidence about how AI performs, which is the subject of Section 2. Commercial and open-source tools are named as examples of a kind of tool, and naming them is not an assessment or an endorsement.

AI's contribution is easiest to understand by following the questions an audit team is trying to answer. Research gives the team possible explanations to investigate. Program records show what has happened in a particular setting, while interviews and further analysis help explain the differences between expectations and results. What auditors learn can change which records they seek, whom they interview and how they interpret an earlier observation. AI supports this developing inquiry by making dispersed information easier to examine and by helping the team pursue connections across sources.

The example is a performance audit of a national affordable-housing fund. The team wants to know whether homes are being completed on schedule, whether they serve the needs the fund was intended to meet, whether the intended households can use them and whether affordability can be sustained. Following these questions from research to reporting shows how several kinds of tools work together and where auditors exercise judgment as the investigation develops.

The audit institution authorizes fuller AI assistance for this audit, in order to learn whether it supports broader comparisons and better-supported explanations within the team's resources. The methodology lead sets the conditions, which include approved arrangements for confidential information, a manager responsible for each use and a record of significant instructions, outputs, corrections and effort. The engagement director remains accountable for the audit's conclusions. At the end, the audit institution assesses what the assistance contributed and which uses deserve to continue.

The tools available range from open-source software that an audit institution can adapt to commercial applications and managed services. Some provide one function, while others combine research, retrieval, analysis and drafting. An institution's choice depends on its information requirements, its existing systems and its capacity to configure and support the arrangement, including where records are processed and what the service terms permit.

A.1 The program and the audit questions

The fund supports non-profit providers and private developers through agreements that set completion dates and rent conditions for a specified period, with payments released at milestones. Funding guidance gives priority to locations and housing types identified in a needs assessment, including larger and accessible homes, while also considering readiness and delivery capacity.

These objectives call for several kinds of evidence. Completion records help establish progress against schedule, and project locations and characteristics show how provision relates to need. Allocation files and household accounts help explain access and suitability, while rent records and operating plans inform judgments about current compliance and continuing affordability. Connecting these sources lets the team examine whether the program is delivering homes that the intended households can use and continue to afford.

The program's progress report leads with 8,400 homes delivered, alongside measures of funding and commitments, although neither the report nor the department's logic model defines "delivered". Exhibit A.1 gives the figures that the team establishes during planning, using occupancy permits as the measure of completion.

Exhibit A.1. Delivery figures established during planning

Region	Due for completion by period end	Completed	Reported as "delivered"
A	3,000	2,900	2,900
B	2,500	2,000	3,500
C	2,000	800	2,000
Total	7,500	5,700	8,400

Note: Completions are measured by occupancy permits. The figures are constructed for this guide.

Completions stand at 76% of the homes due nationally and 40% in Region C. Officials explain the delays mainly through costs and the capacity of providers, while providers emphasize municipal approvals. These accounts give the team possible explanations to investigate. The wider questions also remain important, since the completion total alone says little about which needs the homes serve or the circumstances of their occupants. Exhibit A.2 summarizes the assistance used at each stage.

Exhibit A.2. AI assistance across the audit

Stage	What the tools help produce	Contribution to the investigation	Judgment that supports its use
Planning	Research searches and source-linked evidence records	Findings compared across studies, with their settings and qualifications	Auditors select sources, define fields and inspect consequential extractions
Planning	A synthesis, an initial brief and an account of program assumptions	Questions formed, definitions compared and relationships identified for investigation	The team examines interpretations before they shape the plan
Fieldwork	Matched records, project profiles and interview guides	Revealing cases selected and their records connected to the questions	Auditors assess links, choose cases and review proposed questions
Fieldwork	Transcripts, themes and comparisons of offers, homes and rent terms	Individual experiences compared and differences across cases investigated	The team preserves context and checks the evidence behind significant conclusions
Analysis	Evidence matrices, timelines and rival explanations	Predictions assessed, missing comparisons identified and causes investigated	Auditors obtain further evidence and judge which explanation is supported
Analysis	Comparisons of operating assumptions and suggested sensitivity questions	Pressures on continuing affordability examined	Analysts calculate scenarios on explicit assumptions, and auditors assess their significance
Reporting	Draft findings, consistency checks, summaries and translation	The analysis explained for different readers, with its evidential basis available	Reviewers examine meaning, figures, scope and language

Note: Constructed for this guide.

A.2 Building a basis for the inquiry

Planning determines which questions the audit will pursue and what evidence it needs. AI helps the team build a resource that supports those choices and remains useful as the questions develop. Research, program documents and early observations are then available for repeated examination, with their sources and qualifications identifiable. This supports the planning for sufficient and appropriate evidence described in ISSAI 3000 (INTOSAI, 2019b, ¶101–103).

Creating and using the evidence database

The team begins by examining what is already known about housing delivery, need and affordability. AI-assisted research tools can find studies, help screen them against inclusion criteria and extract information for comparison. Commercial research services such as Elicit combine several of these functions. Used alongside conventional searches and subject knowledge, they help auditors identify relevant work and decide which studies need close reading. The collection includes evaluations, earlier audits and guidance, as well as web pages, spreadsheets, scanned correspondence and early program documents.

Document-processing tools make the contents of these sources available for further work. They recognize text in scans, identify headings and recover the structure of tables, which preserves relationships that would be lost in an undifferentiated block of text. Docling, an open-source document-processing toolkit, and the commercial

Azure Document Intelligence service are examples. The team retains the originals, together with links from extracted passages and figures to their locations, so that auditors can examine the material in context.

The next step turns readable material into evidence records that are useful to this inquiry. A language model extracts information into fields that the auditors define and classifies it by question or subject. For a study, the fields include its setting, method, main finding and acknowledged limitations. One evaluation, for example, discusses the difficulties that small providers face in preparing complete applications. Its record preserves the finding, the program studied and the qualifications that affect its relevance elsewhere, each connected to the supporting passage. Auditors can then compare what studies say about capacity, costs or approvals without losing the differences between their settings, and checking the important extractions keeps those differences visible in later analysis.

Held together, these records form an evidence database that the team can question repeatedly (Section 2.1). Each record retains its source, date and version. Search combines exact words with matching by meaning, so that a question about provider capacity can locate accounts of applications returned for missing information even where the source never uses the word "capacity". Filters focus the results on a jurisdiction, a period or a kind of program. The benefit for the investigation is that the team can pursue a new question across a collection that is already organized and return to the original evidence behind an apparent connection.

There are several ways to supply that evidence to a generative model. Auditors can search and select passages themselves. In retrieval-augmented generation, software retrieves material in response to a question and supplies it to the model to help generate an answer. The team can also review and adjust the retrieved selection before generation. These arrangements can be combined, with different degrees of automation and intervention by auditors. Recording the passages and source versions supplied for each response lets the team examine what the model received and assess how the answer used it.

Generative assistants and source-based research workspaces then help compare findings and prepare a synthesis from the selected material. A general-purpose assistant such as ChatGPT is an example of a tool that reads supplied files and drafts analysis. The important design choice for this audit is how the synthesis remains connected to its evidence. Auditors inspect the selected passages, add a missing study or remove irrelevant material before requesting a revised draft, as in the public-audit prototype described in Box 2.1 (CSIRO, n.d.).

The initial synthesis gives most weight to provider capacity. Because the team can see the sources behind that emphasis, it can also see the reason for it, which is that the collection is dominated by studies of small providers. Capacity therefore becomes one explanation to investigate, alongside costs and municipal approvals. As the audit develops, the team broadens the research and reconsiders how it applies to

this fund, preserving earlier versions so that changes in its understanding remain traceable.

Testing a planning brief

A generative assistant brings the early research and program evidence together in a brief that proposes questions, cases and further enquiries. Such a brief helps the team consider more material when choosing where to concentrate. Its interpretation matters, because an apparently small mistake at this point could direct later work towards the wrong question.

Before relying on these briefs, the manager tests one against a brief prepared by an experienced auditor. Both receive the same early package, which consists of the research synthesis, progress and provider reports, permit and rent records and a sample of agreements. Exhibit A.3 gives the full text of both briefs alongside the scoring summarized in Section 3 (Exhibit 3.2).

Exhibit A.3. Two illustrative delivery-planning briefs

Brief A, drafted by the AI assistant.	Brief B, written by an auditor.
The fund has delivered 8,400 affordable homes, with every region at or ahead of schedule apart from a small shortfall in A [progress and provider reports]. Research identifies costs, permits and capacity as barriers [synthesis]. Region C providers cite permits [correspondence]. Rent confirmations are incomplete in B [rent records]. Obtain C's permit timelines and cost reports and B's missing confirmations.	Permits record about 5,700 completions, below the 8,400 reported as delivered [permits; progress report]. In the agreements read, B and C use delivery for construction start [agreements]. In C, permits cover well under half the homes reported as delivered [permits; provider reports]. Compare completions with agreement schedules, by region, and resolve B's missing rent confirmations before concluding.

Dimension, scored 0–2	Brief A	Brief B
Sources cited beside statements	2	2
Breadth of material used	2	1
Interpretation essential to the delivery-planning decision	0	2
Specific follow-up	2	1
Clear organization	2	1
Total	8	7

Note: The briefs and the scoring are constructed for this guide.

Brief A uses a wider range of material and proposes useful follow-up. However, it treats reported delivery as completion against schedule. Checking that its figures appear in the cited records would not catch this, because the error lies in what the brief concludes from them. The scoring records the failure with a zero, and yet the total still ranks Brief A first. The test shows that a material failure in the interpretation on which a decision depends cannot be offset by strengths elsewhere in an aggregate score. Review therefore has to examine reasoning and definitions as

well as the relevance, validity and reliability of the evidence (INTOSAI, 2019b, ¶106–108).

The manager uses the test to set a condition for planning outputs: a brief that contains a material error of interpretation is corrected before use, whatever its total score. From this point every figure carries its definition, and the team keeps Brief A's useful enquiries into Region C while correcting its account of delivery. The team thereby keeps what the broad brief contributed and directs review to the judgments on which the plan depends.

Preparing records for comparison

The extraction used to organize research also helps auditors compare the program's own records. These include about 300 agreements written on several templates, reports from 150 providers and municipal permit records that use different project identifiers. Reading these documents again for each new question would make comparison cumbersome. Extracting the relevant terms and figures into common fields lets the team examine the agreements together and return to the wording behind any result.

Auditors define fields for delivery, scheduled completion and rent provisions, and a language model returns candidate values with their source passages. Matching methods then connect the records that describe the same project, using exact identifiers where they exist and suggesting other links from names, addresses and providers. The team checks a sample, the uncertain links and some matches that look straightforward. While reviewing Region C files manually, an auditor also identifies two relevant agreement addenda that queries against the evidence database had missed; the team adds them by hand and records the omission for post-audit review. An analyst then uses the prepared records to classify homes by stage and to calculate totals through explicit rules. This division of work makes both the extracted interpretations and the calculations available for examination.

The comparison across all agreements establishes what Brief B's sample had suggested. Region A uses "delivered" for completed homes, while Regions B and C use it for the payment milestone at the start of construction. Payments at the start of construction may be legitimate. The reporting problem is that the national headline adds figures that represent different achievements. By making these definitions visible across the collection, the arrangement helps the team investigate how the headline was assembled and what progress the department can support. With assisted extraction, a comparison across every agreement is practical within the audit's resources, where a review by hand would have covered a sample.

The prepared records remain useful for questions about need, housing characteristics and rent conditions. Natural-language query tools translate a question into a database instruction, and conventional software then selects the records and calculates the results, as in the National Research Council Canada prototype described in Box 2.1 (Liu et al., 2025). An analyst checks what the query

selected and how it interpreted the fields before the result is used. The team can then explore new questions without rebuilding the underlying comparison each time.

Examining how the program is expected to work

AI also helps the team examine the department's account of how the fund is expected to achieve its objectives. A program's logic model describes how its activities are intended to lead to results, together with the conditions on which those results depend. The department's model gives the auditors a starting point. Drawing on the funding guidance, the research and the early records in the evidence database, the AI assistant helps identify assumptions that the model leaves unstated, differences between the documents and relationships that deserve investigation. This gives the team a clearer understanding of what has to happen for the program to succeed and where difficulties could arise.

The expected completion dates depend on providers preparing complete applications promptly and on municipalities approving them within an assumed period. The assistant makes those dependencies explicit and compares them with the research and the early accounts of delay. An explanation centred on provider capacity would lead auditors to examine how applications are prepared, while one centred on municipal processing would direct attention to what happens after a complete application is submitted. The team therefore seeks records that distinguish the two explanations, including application histories, requests for missing information and approval dates. In this way, AI helps turn a broad concern about late construction into specific questions about where delays occur and what might explain them.

The resulting account is a working theory of change, which sets out how the program is expected to produce its results and which relationships still need testing. Auditors examine it against operational experience and further evidence, exercising the judgment that remains central to developing a credible causal account (HM Treasury and Evaluation Task Force, 2026). The team draws its audit criteria from the program's commitments and from appropriate standards. By examining the program's logic, it can investigate why performance fell short and base its recommendations on that explanation.

A.3 Following questions through fieldwork

The organized records help the team choose cases that shed light on its questions. Project profiles combine shortfalls, provider characteristics and reported reasons for delay, and an AI assistant drafts interview guides from those profiles and from the gaps in the evidence. Auditors choose the cases and review the guides, including for any question that assumes the explanation under investigation. For the delivery question, the team selects 40 delayed projects in Region C, covering about 900 of the region's 1,200 homes behind schedule, together with eight projects that are on schedule, for contrast. The access question calls for a different comparison, which includes providers with high and low rates of declined offers.

From interviews to evidence

The interviews give the team accounts of the delays from people involved in different parts of the process. AI helps auditors compare those accounts across projects, drawing attention to recurring experiences and to differences that deserve investigation. Transcription first turns the conversations into searchable text. Speech-recognition tools include open-source options such as Whisper and the transcription functions within commercial meeting services. With the interviewees' consent, the team uses an approved arrangement and retains the links between transcripts, recordings and the projects discussed. Auditors can then examine related passages together while keeping each statement connected to the person and the circumstances behind it.

Qualitative-analysis tools help organize this material under categories that the auditors define. AI features, such as the coding suggestions and summaries offered within commercial software including NVivo, support that work. Providers describe applications sent back, missing drawings and repeated requests for information. Brought together, these passages show a recurring difficulty in getting applications ready for a decision, although providers describe it in different words. Auditors read the proposed groupings, examine what each account contributes and correct categories that split one issue between labels such as approvals "on hold" and "pending".

This pattern gives the team a reason to examine municipal records for the projects concerned. Because municipalities fall outside a federal audit office's direct statutory mandate, the team gathers project application histories through agreement reporting records and voluntary cooperation with the two municipalities. Application histories can show when documents were submitted, whether they were returned as incomplete and how long a decision took once the required information was available. If the applications of inexperienced providers were returned more often, auditors could investigate difficulties in preparing them alongside the time municipalities took to process complete applications. Follow-up interviews explore what lies behind those differences. AI has helped turn separate accounts of delay into a focused investigation of where the process breaks down.

Following the access question

The access question requires the team to understand whether homes suit the households they were intended to serve. Income eligibility establishes one part of that question, although it does not describe the fit between a household's requirements and the features of a particular home. Because the needs assessment emphasizes larger and accessible homes, the auditors examine both the allocation rules and the circumstances of offers that households declined.

AI-assisted extraction helps prepare that comparison from allocation files, offer records, waiting lists and vacancy records. Descriptions of household requirements appear in notes or correspondence, while home characteristics and offer dates are held elsewhere. Bringing the relevant information into agreed fields lets the team set each declined offer beside the household's requirements, the homes available at the time and the home actually offered. Under the institution's privacy and security

conditions, health and disability information is processed only through de-identified extracts inside an approved environment that prohibits external model training or retention. Auditors examine the consequential interpretations and the links between records.

The tenancy files examined show that housed households met the income rules, which is a positive result. The selected providers also recorded 310 declined offers, and reasons can be recovered from correspondence or interviews for 190 of them. Thematic tools help organize explanations involving stairs, bedrooms or location, while links to the original accounts let auditors consider each reason in context. The 120 offers without a recoverable reason remain a limit on the explanation. The comparison shows more than how often households complained, since it helps the team investigate whether suitable homes were unavailable or whether the offer process contributed to the difficulties.

Need, rents and continuing affordability

The same preparation helps compare provision with the needs assessment. In some priority localities, standard homes progressed faster than larger and accessible homes. Retrieval brings the relevant applications and selection reasons into view, while conventional maps and comparisons show the distribution. Auditors then examine readiness, sites, costs and the pipeline of applications before judging how the department demonstrated alignment with its priorities.

Rent comparisons connect occupancy and payment records with the terms of each agreement. Region B has 1,900 occupied homes but rent confirmations for only 1,200. Linking the other 700 records to the relevant provisions helps the team examine the unconfirmed homes systematically. Of these, 620 comply and 80, at one provider, are let above the agreed rent. Checking the links, some records that were not flagged and every record supporting the breach establishes the distinction between a gap in monitoring and actual non-compliance. AI's contribution here is to assist preparation and interpretation across differently worded records, while the comparison itself is carried out through conventional analytical software.

Continuing affordability requires a forward view. Tools extract rents, operating costs, maintenance and reserves from providers' plans into a comparison by period. An AI assistant helps identify changed or missing assumptions and suggests sensitivity questions. An analyst then calculates what happens under explicit scenarios, such as higher maintenance costs or slower rental income. Some providers have updated plans and financial buffers, while others rely on outdated assumptions. The team investigates how the department recognized those different pressures and whether it had a basis for responding before current affordability was affected.

A.4 Developing and testing findings

As evidence accumulates, the team needs to decide what the sources establish together. An evidence matrix connects each question and criterion with observations, supporting passages, contrary material and unresolved issues. The AI

assistant helps populate it, proposes rival explanations and identifies what each would predict. This gives auditors an organized basis for deciding what further comparison is needed and how strongly an explanation can be supported.

Developing the delivery explanation

Timelines help the team investigate the competing accounts of delay. Extraction tools draw application, approval and construction dates from municipal records and provider reports. Analytical software orders the events and calculates the intervals, while the AI assistant helps compare those sequences with the proposed explanations. If provider capacity is the cause, applications should be incomplete or late. If municipal processing is the cause, there should be long waits after a complete application. Auditors investigate whether those intervals actually delayed construction and what other constraints affected the same projects.

The timelines show both patterns, in different proportions. In 19 delayed projects, covering about 450 homes, complete applications were filed on time and approval took a median of 14 months, against the six months assumed in the fund's schedules. Each agreement required approval before construction could start, so a later approval meant a later start. In seven more projects, covering about 200 homes, incomplete applications were combined with long processing. In the remaining 14 projects, covering about 250 homes, the delay arose mainly after approval or the application records were too incomplete to classify. The long approvals affect experienced providers as well as new ones, and they are concentrated in the two largest municipalities.

Asked which comparison is missing, the AI assistant proposes one that the team had not made, namely approval times for similar projects outside the fund. The team obtains those comparator times from published municipal permit data and voluntary disclosures from the two municipalities. Comparable projects outside the fund waited a median of 13 months for approval. The long waits were therefore usual in those municipalities, and the fund's schedules had assumed less than half that time.

This evidence changes the explanation that the team is developing. Provider capacity remains relevant, particularly to the incomplete applications, while approval times that the schedules did not allow for explain the delay in the largest group of projects examined. The manager requires further work on the scheduling assumption before accepting the account. The team also adds research on approval times and updates the synthesis, including its coverage and its assessment of relevance. Findings from other programs remain identifiable in their original settings. Earlier interpretations are marked as superseded, and the working papers and drafts that relied on them are revised and checked against the expanded evidence.

Root-cause analysis extends the question from where the delay occurred to why the program's arrangements failed to anticipate or address it. Planning records show that the six-month approval period came from Region A's faster experience, and officials confirm that Region C's municipalities were not consulted. AI helps organize the chain of questions and identify the links that need corroboration. Auditors can

then separate two kinds of cause. Obstacles to delivery could be addressed through support with applications and earlier engagement with municipalities, while weaknesses in scheduling could be addressed through schedules informed by local conditions. Recommendations follow the causes that the evidence supports, consistent with ISSAI 3000 (INTOSAI, 2019b, ¶127). The findings concern the selected projects, covering about 900 delayed homes, and are not projected onto the rest of the region.

Explaining access and the wider results

The comparison of offers helps distinguish an explanation based on supply from one based on allocation. A supply problem arises where no suitable home was available for the household. An allocation explanation requires evidence that a suitable home was available and could have been offered under the applicable rules and priorities. The examined records point mainly to supply, with allocation contributing at a few providers. Auditors consider the 120 unexplained offers when judging how far that account can extend. They conclude that income rules were met in the examined files, that some intended households could not use the homes offered and that routine reporting did not explain those experiences. Whether this reflects the slower progress of larger and accessible homes remains a question for further examination unless the audit establishes the link.

On alignment with need, readiness influenced selection, but the expected contribution to identified shortages was not assessed consistently. The finding therefore addresses how the department demonstrated that alignment, taking account of feasible sites, costs and available projects. Current rent breaches remain separate from the risks suggested by future operating plans. A lack of consistent monitoring of those plans supports a finding about the department's forward view, without predicting that a particular provider will breach its agreement.

Across these questions, AI helps keep evidence, alternatives and gaps available for comparison. The auditors' explanation develops as they obtain and test evidence, including material that challenges their initial view. The arrangement is useful where it helps auditors examine more sources and keep a clear record of why their conclusions changed.

A.5 Preparing the report

Reporting requires the team to turn its analysis into an account that readers can understand and the evidence can support. Generative drafting tools work from the agreed findings, the evidence matrix and selected passages, with instructions about audience, terminology and source links. The team uses this assistance to organize the findings on delivery, need, access and affordability, making their relationships clear while retaining their different evidential bases. Reviewers compare each proposed formulation with the underlying analysis.

AI-assisted consistency checks also help locate differences between drafts, summaries and verified figures. One summary gives Region C's completions as 900 instead of 800, having borrowed the number of delayed homes in the projects

examined, and a comparison with the checked data exposes the error. The same summary turns income-eligible occupancy into an assurance that priority needs were met. Recognizing that change requires attention to meaning as well as to numbers. Review against the evidence matrix restores the distinction between who met an eligibility rule and whether the available homes served the needs identified for the program.

Summarization and translation make the findings available to different audiences. Machine-translation tools, including commercial services such as DeepL, can apply glossaries to support consistent terminology in a draft. Bilingual reviewers assess the meaning, and reviewers of the public summary check that it preserves significant qualifications. The department's comments may prompt further examination or changes, and the reasons for important decisions are retained in the working papers (INTOSAI, 2019b, ¶87, ¶129–130). AI assists the team in communicating an analysis it has already agreed, and that analysis stays available for review.

A.6 What the audit institution decides

The audit institution finishes with a record of what each use contributed, what checking it required and where corrections were needed. A quality reviewer outside the team examines that record as input to the methodology lead's decision. The engagement director's responsibility for the audit's conclusions stays separate from the decision about future AI use. Exhibit A.4 summarizes the experience that the audit institution considers.

Exhibit A.4. Assessing the contribution and deciding what to continue

Use	Contribution	Effort or limitation to weigh	Decision
Research and evidence records	A reusable collection with findings, settings and qualifications available for comparison	Source selection, field design and checking consequential extractions	Continue, retaining source links and identifiable versions
Retrieval and synthesis	Questions pursued across the collection, and research reconsidered as the audit develops	Missed relevant records, and review of the selected material	Continue for investigation; test retrieval before expanding reliance on answers
Planning briefs	A broad proposal with useful follow-up	A critical error of interpretation concealed by the aggregate score	Require correction of material errors before use, whatever the total score
Extraction, matching and queries	Definitions and rent terms compared across connected records	Preparing fields, checking links and verifying what queries select	Continue where the scope gained justifies preparation and review
Interviews and access comparisons	Individual accounts and offers connected to records and circumstances	Consent, protected processing, review of classifications and missing reasons	Continue, with context and evidential limits retained
Explanations and scenarios	Rival accounts examined through timelines and explicit assumptions	Corroboration, further evidence and analyst review	Continue as support for inquiry and judgment
Reporting assistance	Drafts and summaries tied to the agreed analysis	Corrections of figures and meaning, translation and final review	Continue, with review against the evidence and intended scope

Note: Constructed for this guide.

The assessment separates two kinds of effort. Defining fields and resolving project links is preparation that supports several later uses within the audit, whereas checking a misleading interpretation each time it appears is a recurring demand on professional attention. The audit institution considers whether the comparisons and questions pursued justified both. Broader coverage or clearer reasoning can matter even where time savings are small, provided the resulting work can be supported.

The assessment also compares the arrangement with credible alternatives. One is conventional searches, databases, spreadsheets and targeted or complete review, combined with the auditors' own analysis. Another is a narrower AI arrangement that assists extraction and retrieval and leaves the development of explanations and the drafting to the team. Both could connect evidence and test explanations. The relevant comparison is what additional assistance the fuller arrangement supplied, how well it performed and what it cost in preparation and review. This constructed audit illustrates contributions that an institution could examine, but it supplies no measured comparison between the arrangements.

On this experience, the methodology lead continues the useful preparation, comparison and drafting as controlled uses, authorized audit by audit, and the

condition on planning briefs becomes a standing requirement. During the audit, several answers drawn from the evidence database also missed relevant documents that auditors found by other routes. Before any finding relies on such answers, the audit institution will therefore test retrieval with questions whose relevant records are already known, including contrary material, and examine the misses. Until then, those answers serve as leads for auditors to examine.

The practical lesson of the example is that a fuller use of AI can be designed around the continuing work of investigation. The evidence database, the prepared records and the linked analyses help the team carry questions across stages without assembling the same material repeatedly. Auditors use that assistance to examine alternatives, obtain further evidence and explain their judgments, and leaders assess the whole contribution and allow experience to change the arrangement. Sections 4 and 5 examine the institutional safeguards, continuing support and professional capacity needed to sustain this work across an audit office. Section 6 then synthesizes what these requirements mean for the profession, arranges the guide's six recommended actions into a practical first-year sequence and looks ahead at how public-sector audit could evolve.

Production note

This guide was researched and produced through an AI-assisted evidence-synthesis workflow anchored in a local, source-linked SQLite database rather than drawing claims from a language model's memory. The collection brings together published research and field experiments, performance audits of government agencies, Canadian and international assurance standards, regulatory guidance, public-audit AI strategies, early prototypes and industry surveys.

Before analysis began, documents were gathered through structured and gap-directed searches, deduplicated across file versions and pinned to local copies with citation details. PDFs and web reports were converted using structured document-parsing tools into numbered, page-bound text passages while keeping original files and page images linked underneath. Because automated extraction can scramble complex layouts, tables, sidebars or rotated margin watermarks, extracted text was treated as a checkable derivative and original pages were reopened whenever a quotation, figure or layout looked uncertain.

The resulting database holds 153 catalogued entries, 106 full source profiles summarizing each document's methods and limits, 406 structured research records and 7,688 indexed text passages tied back to specific pages. To keep evidence distinct from interpretation, the database stores three layers separately: what an individual source reports, how sources compare across a topic and what a finding might mean for audit leaders. After checking full-text access, narrowing to the guide's scope and placing unresolved items on hold, 65 verified sources entered the final reference list.

AI tools assisted across both research and production. During research, models summarized sources neutrally before introducing the guide's argument, and ran omission checks after early two-model tests showed different readers surfacing different passages from the same papers. During writing and production, AI tools helped draft working text from structured evidence notes, verify citations and worked-example arithmetic, run a separate-model challenge on the executive reading path, generate the three conceptual diagrams and compile the Word and PDF editions from a single Markdown master.

In line with the guide's own principles, checking focused on load-bearing claims, quotations, held sources and framing, and separate-model review served as a diagnostic check rather than independent peer review. The author stayed actively engaged throughout writing and editing, shaping the structure and argument, directing revisions across each section, reworking prose and tone directly and taking responsibility for all final conclusions.