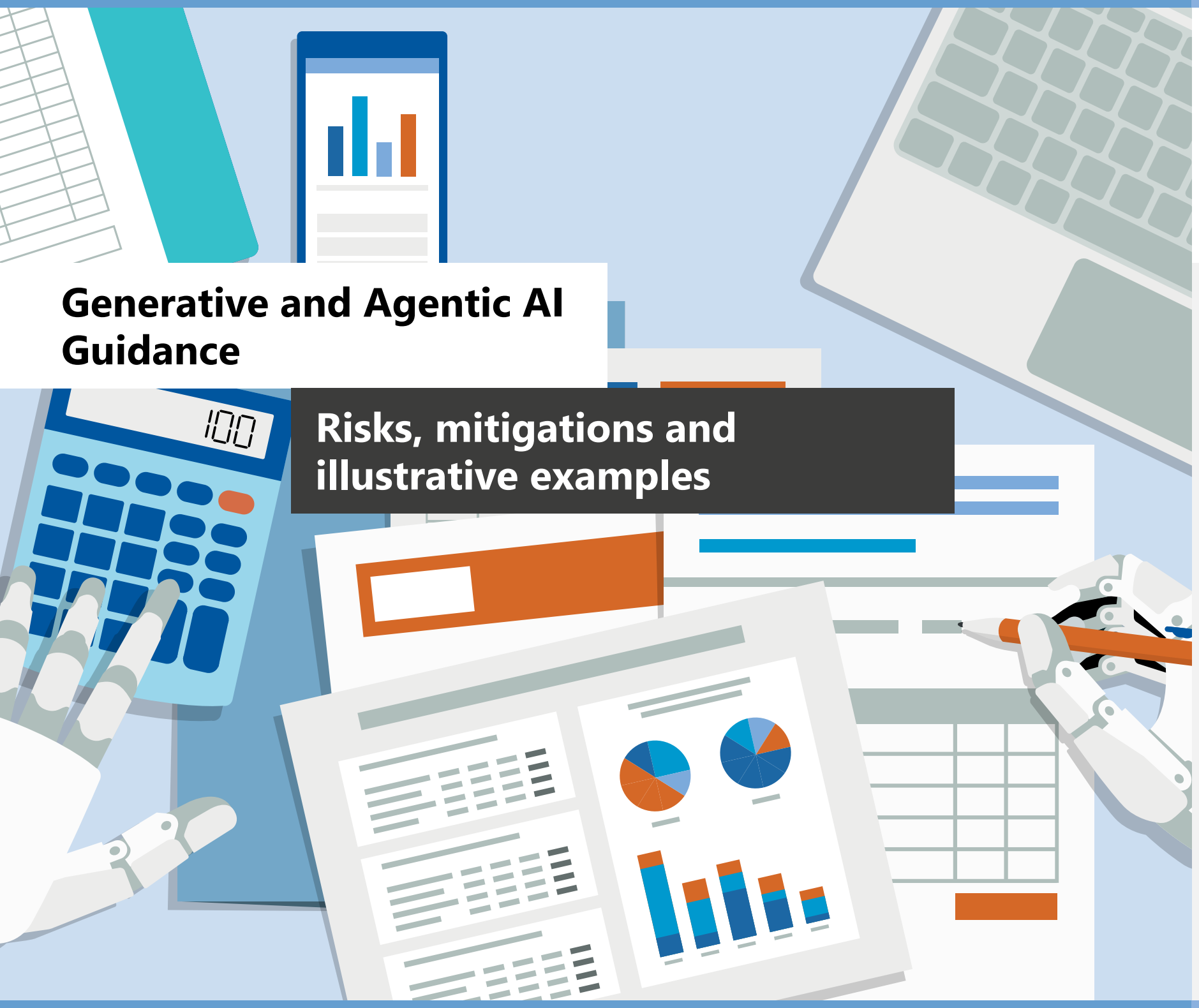
Generative and Agentic AI Guidance

Risks, mitigations and illustrative examples



Financial Reporting Council

March 2026

Financial Reporting Council
13th Floor
1 Harbour Exchange Square
London
E14 9GE

Contents

Introduction

The FRC supports innovation and the appropriate use of artificial intelligence (AI) to promote high audit quality, growth in the UK economy and the public interest. In particular, generative and agentic AI tools have the potential to significantly enhance audit quality across a diverse set of audit procedures and activities.

However, these technologies pose risks to audit quality too, relating to the risk of deficient outputs, the risk that outputs are misused and the risk that the audit methodology is not compliant with auditing standards. This guidance discusses those risks and how they may be mitigated through appropriate system design and development, certification, staff education and governance and human in the loop – see below for definition – review and oversight. It further includes illustrative examples that portray the consideration of risks and mitigations in the context of potential use cases. This guidance is aimed at the central technical teams within audit firms who are responsible for the development of generative and/or agentic AI tools and supporting methodology.

Definitions

Generative AI (GenAI) systems are artificial intelligence technologies that can generate content in response to prompts, using learned patterns from large training datasets. **Large language models (LLMs)** are a specific form of GenAI that consume, interpret and generate text. For the purposes of this guidance, reflecting the current audit landscape, references to GenAI models or systems should be understood as referring to large language models or systems comprising large language models and non-AI components.

Agentic AI systems are artificial intelligence systems that can orchestrate and execute multiple components and/or tasks toward a goal, with some degree of autonomy. These systems typically include one or more LLMs that serve as both the orchestrating "brain" and as part of the execution layer, performing functions such as researching, reasoning, generating content or critiquing outputs.

Both non-agentic and agentic AI systems can include one or more LLMs, as well as components that support data acquisition, computation or interfacing with other systems. The distinguishing feature of agentic systems is that the system has the agency to pursue a goal independently across multiple steps, potentially with some degree of autonomy to determine how to do so, without step-by-step human instruction.

In the context of an AI system, a **human in the loop** is a human who directs, authorises or reviews the system's actions or outputs at runtime. In a non-agentic system, each action is individually prompted and reviewed as appropriate. In an agentic system, the human in the loop may set the initial goal, if this is not fixed by the system designers, provide information, and oversee or review actions or outputs at specific control points.

Risks

Risks

Introduction

The use of generative or agentic AI on an audit engagement may present risk across many different dimensions. For example, there could be organisational risk, if human knowledge and expertise are eroded over time; cybersecurity risk, if audited entity data is input into cloud-based AI applications; or financial risk, if the cost to develop and run the technology exceeds the returns. In this guidance, however, we limit our attention to risks to audit quality on engagements where it is deployed.

Risks to audit quality from the use of an AI tool fall into three categories:

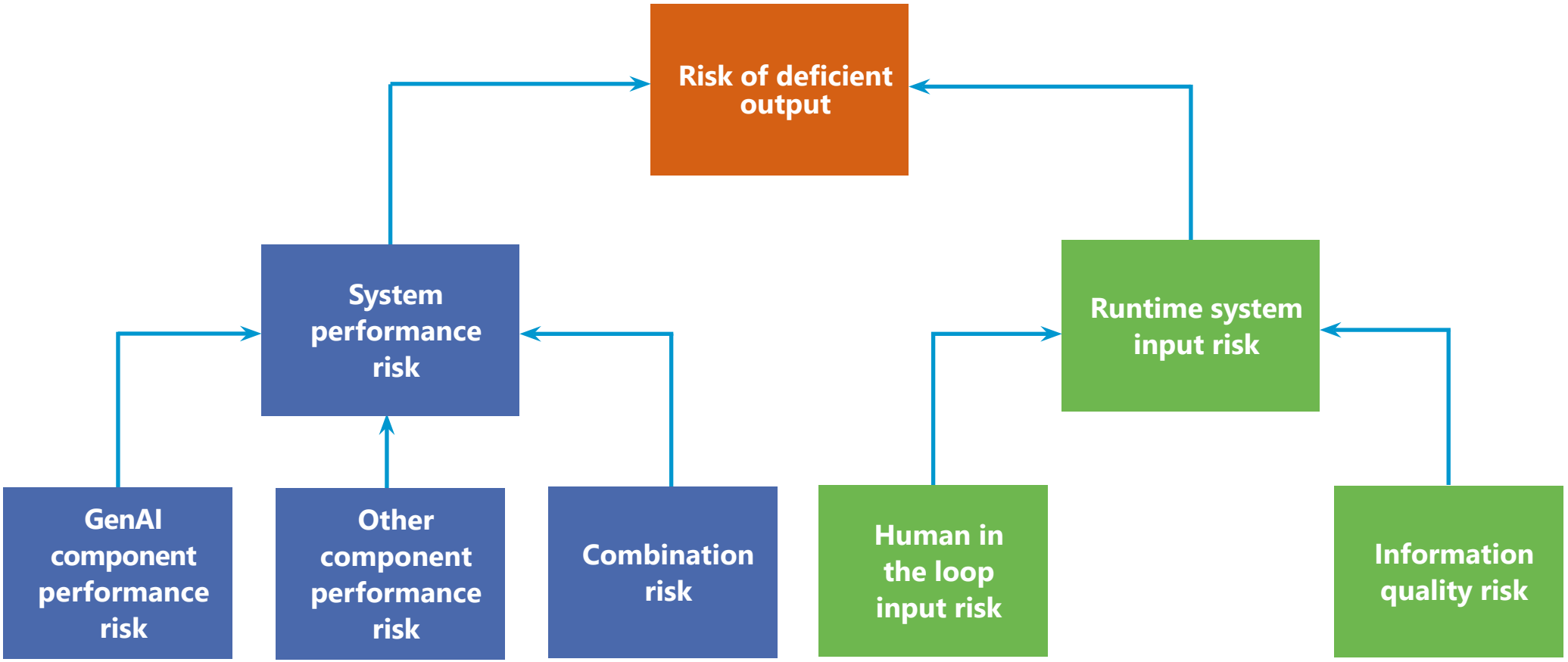
- Risk of deficient output – the risk that the output of the AI tool is deficient;
- Risk of misuse of output – the risk that the output is misused, despite the output itself being appropriate;
- Risk of non-compliant methodology – the risk that the firm's methodology permits approaches, which include the use of the AI tool, that fail to meet auditing standards, even where the output of the tool is appropriate and is used as specified in the methodology.

We will explore each of these in turn.

Risk of deficient output

Introduction

The output of an AI tool may be deficient due to an issue with the performance of the system itself, or due to issues with inputs to the system at runtime. System performance issues can arise from performance issues with components, be they GenAI components or not, as well as error or distortion that emerges from how components combine with each other. **For AI systems that are comprised of a solitary LLM, system performance reduces to GenAI component performance.** Inputs to the system at runtime, either from humans in the loop or any information sources accessed, can also create risks of deficient outputs.



Risk of deficient output

System performance risk – GenAI component performance risk

Contemporary LLMs can exhibit several performance issues that can lead to deficiencies in their outputs, which can in turn pose issues for system performance and audit quality. These LLM output deficiencies fall into five categories:

- **Hallucinations** – information that has been fabricated by the LLM;
- **Omissions** – the absence of information that should be included;
- **Distortions** – misrepresentations of the meaning, emphasis or implications of information;
- **Faulty reasoning** – unsupported or illogical arguments or conclusions;
- **Inconsistencies** – information that is internally inconsistent, or inconsistent with prior outputs without justification.

These deficiencies ultimately arise due to four fundamental limitations of LLMs, as well as the finite compute they can access:

Limitation	Description
Lack of semantic awareness	LLMs generate text by recognising and reproducing statistical patterns in language, without true understanding of meaning, context or logic. This can mean outputs sound coherent but distort meaning, misinterpret context or apply unsound logic.
Dependence on training data	An LLM's learned knowledge is entirely determined by the data on which it was trained, and it cannot independently ground its claims in real world sources of truth. If the training data are incomplete, biased, inaccurate or outdated, outputs of the model may share those same limitations.
Finite model capacity	The amount and complexity of patterns that an LLM can learn is finite, limited by its number of parameters. This can lead to oversimplifications, inaccuracies, distortions or faulty reasoning in outputs. The information that an LLM can integrate at any point in producing an output is finite too; they therefore prioritise and allocate their attention unevenly, which can lead to information being distorted or lost in the output.
Finite context window	LLMs operate within a finite context window, meaning only a bounded volume of information is available for the model to consider, or condition on, when generating an output. This can lead to important detail being lost or overlooked and therefore omissions, distortions or inconsistencies in the outputs.

Risk of deficient output

System performance risk – GenAI component performance risk

LLMs can play a range of roles depending on how they are prompted. These roles are not inherent properties of the model but explanatory constructs to support articulation and discussion of the ways an LLM may be used. An LLM can play more than one role at any one time, if the prompt instructs it to perform multiple tasks.
Some of the main roles an LLM can play are researcher, reasoner, writer, summariser, critic and orchestrator, with the orchestrator role only found in agentic systems.

Role	Description
Researcher	The LLM retrieves information from available sources to support a task.
Reasoner	The LLM reasons logically about information to form conclusions.
Writer	The LLM generates original content based on a prompt. This may constitute the output of the AI tool, or it may enable the LLM to issue instructions to other components.
Summariser	The LLM summarises text, guided by any context and priorities provided by the prompt.
Critic	The LLM reviews and critiques an input, which may itself be the output of another LLM or component of the system.
Orchestrator	The LLM proposes how actions and components should be coordinated to achieve the specified goal, as it has interpreted it.

In agentic systems, there is often a need for LLMs to play multiple roles. Sometimes this will involve multiple LLMs being present in the system, for example when cheaper models can suffice for some tasks, or different specialisations are required, though even here LLMs may play different roles at different times. In many cases, however, the system will have one LLM that plays multiple roles over the course of producing an output, in response to different prompts.

Risk of deficient output

System performance risk – GenAI component performance risk

The way in which an output deficiency in a GenAI component can manifest into an audit quality issue will depend on the role that the LLM is playing in the system, and the purpose for which the system is being used on an audit. The accompanying illustrative examples discuss some of the risks for a selection of common use cases.

It is often relatively clear how the presence of a component output deficiency like a hallucination or distortion could lead to the system not performing as intended. The orchestrator role is so foundational to agentic systems, however, that we will explore the role itself and related risks in greater detail.

In an agentic system, the orchestrator's responsibilities may include:

- Interpreting the system's goal as specified by the human;
- Designing a work program to achieve this goal, potentially decomposing it into discrete elements if it is complex;
- Integrating the outputs of other components;
- Managing iteration and feedback logic.

The orchestrator may not always be granted all of these responsibilities. For example, if the system is going to be performing the same task each time, the work program may be encoded by the system designers and the orchestrator may simply manage the execution of this. The degree of autonomy granted to the orchestrator is a complex strategic decision for the system designers; increased autonomy for the orchestrator often results in greater flexibility for the system, but may carry greater risk.

Each of these responsibilities involves the generation of outputs that are consumed by other components within the system. The presence of any of the 5 categories of component output deficiencies discussed previously in any of these outputs can create a range of risks to the performance of the system, especially as its outputs often underpin the rest of the system's activity.

To illustrate this, we will select different component output deficiencies for each responsibility and show one way that this can create risks:

Risk of deficient output

System performance risk – GenAI component performance risk

Responsbility	Indicative output deficiencies	Potential risk to system performance
Goal interpretation	Omission	The LLM may produce an output that omits important dimensions of the goal, enhancing the risk that the output is deficient.
Work program design	Faulty reasoning	Faulty reasoning may mean the LLM creates a work program that is not appropriate for producing an output that meets the goal as interpreted in the previous step.
Integration of outputs	Hallucination, faulty reasoning	The LLM may invent conclusions and attribute them to component outputs. This can lead to the final output containing unsupported conclusions. It may assert these conclusions overconfidently, which is a form of faulty reasoning.
Management of iteration and feedback	Distortion, inconsistency	The LLM's understanding of the goal may drift over the course of the system's activity if the meaning of the original prompt gets altered through incremental compression to manage space in the finite context window. This can lead to the LLM producing iteration or feedback instructions that misstate the goal, increasing the risk that the final output does not align with the intent of the human.

Risk of deficient output

System performance risk – Other component performance risk

Both agentic and non-agentic AI systems can incorporate components that are not themselves LLMs, which can present risks to system Both agentic and non-agentic AI systems can incorporate non-LLM components, which can present risks to system performance if they fail to perform appropriately. These components include:

- **Data acquisition components**
  - These obtain, extract or transform data so that it can be used by the rest of the system.
  - Examples: data connectors, document loaders, search engines, optical character recognition engines, data mapping tools, retrieval augmented generation retrieval modules
- **Rules-based processing components**
  - These apply predefined, rules-based logic to produce outputs
  - Examples: workflow engines, prompt management components, transformation scripts, validators, calculators
- P**redictive machine learning components**
  - These apply patterns to forecast, classify, score or detect
  - Examples: anomaly detection components, provision estimation components
- **External action components**
  - These allow the system to connect with other systems to perform actions
  - Examples: connectors to email, record or transaction systems
- **Operational infrastructure**
  - These enable the system to run and coordinate its tasks reliably
  - Examples: task schedulers, authenticators, storage, cloud runtime platforms, message queues

If these components suffer from performance issues, the system may operate with incomplete or erroneous information, process information incorrectly, or simply not run at all, leading to clear risks that the system produces deficient outputs.

Risk of deficient output

System performance risk – Other component performance risk

Prompt management components may be found in some AI systems that are not prompted by the human in the loop at runtime. These components provide LLMs with prompts that are written by the system designers, or constructed from prompt templates and information provided by the human in the loop or the outputs of other components. In agentic systems, this function may be performed by the workflow engine. For the purposes of this guidance, component performance risks include those that arise from design-time configurations or instructions, rather than just execution risks. Specifically, for prompt management components or workflow engines, this includes the risk that any prompts or prompt templates written by the designers are not fit for purpose. This mirrors the risks relating to the specification of the task and the provision of any supporting information by the human in the loop, which are discussed in the later section on human in the loop input risk.

The workflow engine, found in agentic systems, is an especially important component as it coordinates the rest of the system. The workflow engine, often influenced by the outputs of an LLM in the orchestrator role, controls:

- The state that the system is in, for example goal interpretation, task allocation, execution or review;
- Which components are called and in what sequence;
- The prompts that components are supplied with, constructing them from template prompts written by the human designers and other information, including human in the loop system inputs and outputs of other components;
- What tools can be called by LLM components;
- How the system recovers from errors.

Issues with the performance of the workflow engine can present increased risks, as it coordinates the rest of the system's activity.

Risk of deficient output

System performance risk – Combination risk

AI systems can produce deficient outputs due to the combination of components with each other, or with themselves through iterative use of the same component. This risk can be segmented into three main categories: amplification risk, information distortion risk and interface risk.

Amplification risk

This is the risk that individually insignificant errors or biases introduced by one component get magnified when subsequently processed by itself or other components, leading to potentially significant deficiencies in the final output.

Information distortion risk

This is the risk that information can be distorted as it is sequentially processed by components, even where each component performs as it should, to the extent that the output is deficient.

Interface risk

This is the risk that errors or performance deficiencies arise from components not communicating as intended. This could be for a range of reasons, including data transfer failures or semantic mismatches, where components interpret content differently.

If the AI system has components that are tasked with identifying and correcting these dynamics, then these risks could be construed as performance risks for those components.

Risk of deficient output

Runtime system input risk – Human in the loop input risk

Human in the loop inputs can cause system outputs to be deficient for a range of reasons. There are three main categories of human in the loop inputs, each of which poses its own risks:

Task specification and supporting information

These inputs specify the goal of the task, how the system should approach it and any supporting context or information the system should consider. These inputs include prompts, uploads of supporting information and, for systems with a prompt management component or workflow engine, structured inputs through forms or templates.

The quality of these inputs will significantly affect the quality of output from the system. In particular, risks of deficient output are posed by:

- **Incorrect task specification** – the goal or instructions that are communicated to the system are not aligned with the intended purpose, or are inconsistent with each other, potentially leading to outputs that are not fit for purpose.
  - Example: An AI tool is prompted to individually compare a series of controls against specified criteria to evaluate their design and then aggregate the results to form an overall evaluation of the control environment. However, the tool is not prompted to look at shared points of weakness or interdependencies between controls and, as a result, the overall evaluation may overstate the effectiveness of the control environment. Therefore, there is a risk of a deficient output as a consequence of the task specification not being aligned with the intended purpose of evaluating the control environment.
- **Ambiguous task specification** – the task is defined vaguely, leaving too much scope for interpretation by the system which may result in outputs that are inconsistent or not fit for purpose.
  - Example: An AI tool is prompted to summarise a complex document, without any further instructions. The output may compress each part of the document by a similar amount, despite some areas being much more relevant to the audit. This risks important information being lost in the summary.
- **Incomplete, misleading or erroneous supporting information** – there are issues with the supporting context or materials provided to the system, which may cause the output to not be fit for purpose.
  - Example: The auditor uploads a prior year memo as context for a task. The memo includes information about the entity that is no longer accurate, which affects how the AI prioritises certain information and means the final output lacks important considerations.

Risk of deficient output

Runtime system input risk – Human in the loop input risk

Control point decisions

These inputs involve a human in the loop determining whether and how the system proceeds at defined points in its operation. These decisions may include authorising certain actions, including the use of tools; choosing whether the system continues or stops operating or determining what the system does next.

If these decisions are not appropriate, for example if the auditor does not understand the scope or purpose of the task, lacks some relevant technical knowledge or exhibits automation bias, the outputs of the system may be deficient. These inputs are only relevant for agentic systems.

Operational configurations

These inputs are the selection of settings by a human in the loop that determine how certain aspects of the system operate. These may include selecting between reasoning modes; setting time or iteration limits or enabling or disabling tool or data access. If these are not appropriate, the outputs may be deficient.

Risk of deficient output

Runtime system input risk – Information quality risk

AI systems are often enhanced with the ability to access information sources at runtime to provide current and/or task-specific information, beyond that which is found in model training data, to improve the quality of outputs. These sources may include:

- Internal technical or policy documents;
- Prior workpapers;
- Financial or transactional databases, potentially including audited entity data if authorisation obtained;
- Professional standards;
- Externally produced guidance;
- Other external information sources, accessed through application programming interfaces, which can include internet search engines.

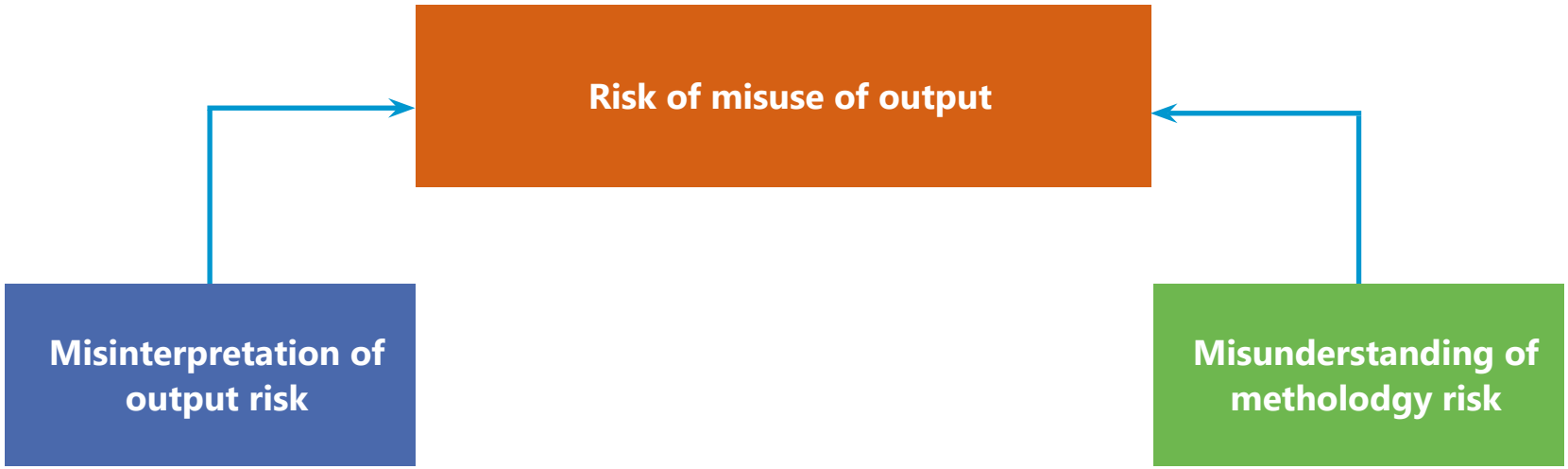
If this information is misleading, erroneous, incomplete or irrelevant, or not effectively catalogued to support consistent retrieval, system outputs may be deficient.

This category of risk deliberately excludes risks arising from issues with training data or information provided by a human in the loop at runtime, as these are included in previous categories.

Risk of misuse of output

Misinterpretation of output risk and Misunderstanding of methodology risk

The output of an AI tool can be misused if a human misinterprets the output of the system or misunderstands how the output should be used in the context of the methodology.



Misinterpretation of output risk

This is the risk that the meaning, scope, limitations or certainty of an output of an AI tool is misinterpreted by the audit team, which can lead to inappropriate conclusions or actions even when the output itself is appropriate. For example, an AI tool may be used to check whether a series of contracts contain leases and summarise those sections if they are present. The auditor may misinterpret the output of this tool as a complete summary of the contracts, and therefore not perform any other procedures to identify, for example, indicators within the contracts of present obligations that may require the recognition of provisions.

Misunderstanding of methodology risk

This is the risk that the auditor does not appropriately understand how use of the AI tool fits into the methodology, which may lead to the output informing inappropriate conclusions or actions. For example, the auditor may not understand that the output of an AI-enabled risk assessment tool was intended to inform their judgement rather than being itself a conclusion.

Risk of non-compliant methodology

Generative and agentic AI enable many new forms of audit procedure, and it requires significant professional judgement to determine how to incorporate these into audits in ways that meet auditing standards. In some cases, they will replace or complement a traditional procedure in a modular manner. In others, they may inspire a significantly different audit approach.

The audit firm's methodology will often guide teams on what would be appropriate with respect to these sorts of judgements, but there is a risk that it permits or recommends approaches that result in audits that fail to meet auditing standards.

This risk is not unique to technology or AI; it arises for any revision of the methodology for a new audit procedure. However, the risk may be increased in the context of AI-enabled procedures, as it can be challenging to compare their outputs to those of traditional audit procedures or calibrate them against requirements of auditing standards.

For example, AI-enabled procedures may produce outputs that directly relate to entire populations, rather than samples, but the insights relating to each individual item within those populations may be less conclusive than they might be in the context of traditional substantive audit work. It may be impossible to compare the persuasiveness of these outputs quantitatively; professional judgement must be exercised to determine what conclusions may be appropriately drawn from the outputs of AI-enabled procedures.

It is not necessary that the outputs of an AI-enabled procedure exactly match the persuasiveness of those of a procedure it replaces, so long as the methodology does not overstate their persuasiveness, and the overall approach meets auditing standards.

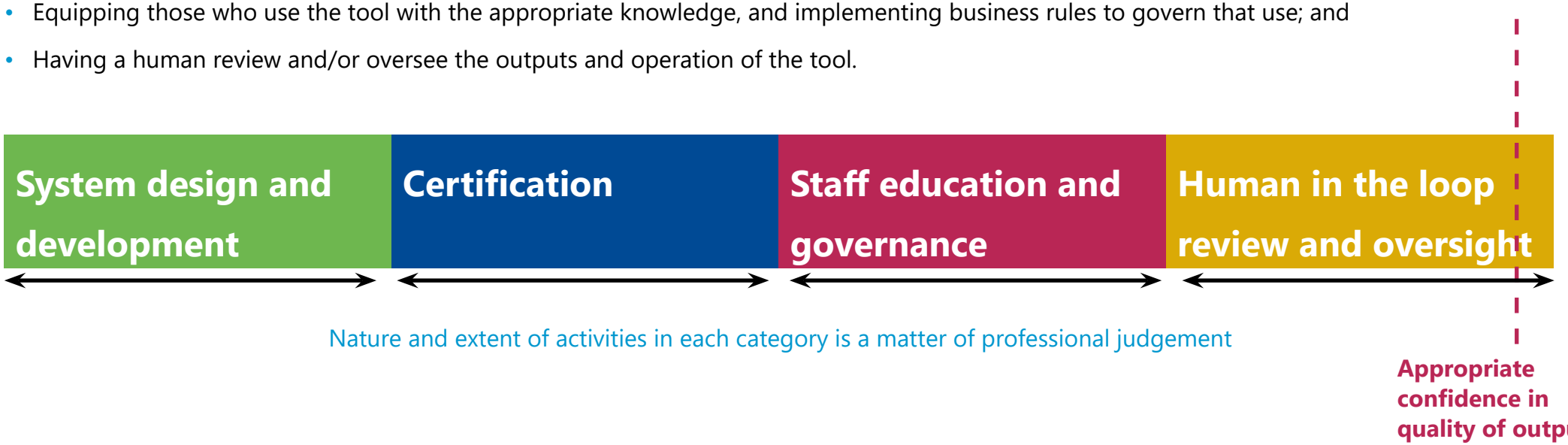
Mitigations

Mitigation of risk of deficient output

Introduction

The risk that the AI system produces a deficient output can be mitigated in a number of ways. These include:

- Designing and developing the system in a way that is responsive to the intended use;
- Putting the tool through a robust certification process;
- Equipping those who use the tool with the appropriate knowledge, and implementing business rules to govern that use; and
- Having a human review and/or oversee the outputs and operation of the tool.



It is a matter of professional judgement for each use case how much confidence in the quality of output that it is appropriate to obtain, as well as how that confidence is obtained. The judgement as to how much confidence in the quality of an output is appropriate to obtain is sensitive to how that output will be used and relied upon. Where the output is to be relied upon as audit evidence, the auditor's evaluation of their confidence in the quality of an output consists of their evaluation of the quantity and quality of evidence obtained, not just its reliability.

The purpose of audit is to reduce audit risk to an acceptably low level. If an AI enabled procedure, presuming no deficiencies in the AI output, may reduce audit risk by a significantly greater amount than the alternatives, then it may be appropriate to tolerate a higher risk of deficient outputs from the AI tool, if the expected audit risk as a result is nevertheless lower than if alternative procedures were performed.

Mitigation of risk of deficient output

Introduction

The nature and extent of mitigating activities performed from each category is a matter of professional judgement, and is sensitive to the activities performed from other categories. For example, if a tool consistently produces high quality outputs in testing, and staff are trained to use the tool appropriately, then it may be appropriate for review activities to be less extensive than if output deficiencies are frequently observed or there is greater uncertainty about how staff may use the tool. Where there is significant scope for the human in the loop to influence how the tool operates through their choices, for example if they write their own prompts, the confidence in output quality it is possible to obtain from centrally testing the performance of the tool may be less than if there is greater central control over how the tool operates. If so, it may be appropriate that the review of the output is more extensive.

Performing some mitigating activities from each category may be appropriate for all tools and use cases, but the efficiency of each form of mitigation may vary significantly across tools and use cases, so the firm may exercise professional judgement to determine how to obtain appropriate confidence.

Mitigation of risk of deficient output

System design and development

Appropriate system design and development plays a key role in mitigating risks of deficient output. System design and development mitigations can be cross-cutting in nature, or targeted at specific categories of risk, such as GenAI component performance risk.

Cross-cutting mitigations

- **Workflow design.** Designing a workflow for an AI system involves consideration of the tasks the system will be set and how it may approach them. For agentic systems that may be set a variety of tasks, this consideration may be at a high level, as much of the detail of how the system operates for each task may be informed by an orchestrator LLM, which may design a detailed work program to be completed by the system. For other systems, the designers may specify a detailed workflow for both humans in the loop and the system itself to follow; this usually mitigates risk to a greater extent, at the cost of flexibility. In either case, the designers consider what components will be required and, at different degrees of granularity, what activity the system must complete to achieve the goal, what roles an LLM may be called upon to perform, how components may combine and what human in the loop control points may be appropriate.

  The design of a workflow that is structurally appropriate for consistently achieving the goal significantly mitigates the risk of deficient system outputs. Some examples to illustrate this are:

Risk	Possible workflow design mitigation
Workflows that require AI systems to perform tasks beyond their capabilities may increase the risk that their outputs are deficient.	The determination of an appropriate perimeter for the workflow, leveraging humans to perform tasks that AI is less adept at.
Fixed, linear workflows may be too brittle for some use cases, as they lack the ability to revisit earlier steps or adjust if presented with new or unexpected information, which can lead to deficient outputs.	The inclusion of conditional looping or branching in the workflow.
Local optimisation at each step may mean the system produces coherent intermediate outputs but never synthesises them into an output that achieves the task.	The inclusion of synthesis steps in the workflow, or a step to test the final output against the initial goal.

Mitigation of risk of deficient output

System design and development

- **Component choice.** Including the appropriate components within the system, in the context of the expected workflow, means components can be allocated tasks that match their capabilities, reducing the risk of component performance issues.
- **Rules-based protocols.** Specifying deterministic criteria and instructions that govern how components behave in certain situations mitigates a range of performance risks for a wide array of components, including risks arising from the combination of components. In particular, rules-based protocols are a useful complement to the inherently probabilistic and not fully predictable behaviour of LLMs. Some common forms of protocol are:
  - **Input schema.** These contain structural or content criteria for each form of input.
  - **Output schema.** These contain rules-based criteria that component outputs must meet. For example, they may specify that outputs must be presented in a certain format; fall within plausible ranges; cite their sources or are numerically internally consistent.
  - **Scope and authority boundary rules.** These are rules that define what actions the system can or cannot perform, including for example when it can iterate; what tools it may use and when human authorisation is required.
  - **Issue response rules.** These govern how the system responds to issues or errors. For example, they may define when issues should be escalated; how the system can fail safely; or what alternative steps it can perform.
- **Ongoing support and version control.** Ongoing support and version control for both the system as a whole and individual components mitigates the risk that issues arise over time that cause deficient outputs. Various components, including AI models, may be updated periodically either by the audit firm or a third party; implementing a process for monitoring this, and more generally controlling if and how updates are deployed to the AI systems that are the subject of this guidance, is an important mitigation of risk

Mitigation of risk of deficient output

System design and development

Mitigations for GenAI component performance risk

- **Workflow design.** Workflow design was mentioned in the previous section as a cross-cutting mitigation for structural issues with the workflow that may lead to system output deficiencies even where each component performs as intended by the designers. In this section, we discuss ways that workflow design can mitigate GenAI component performance risks by compensating for certain limitations of LLMs:
  - **Distribute the cognitive load across steps and/or components.** This can mitigate risk of deficient outputs in many ways:
    - Each time an LLM is prompted, it can deploy a finite amount of compute to produce an output. By splitting tasks into less complex ones, and prompting LLMs to work on each in turn, the system can deploy more compute overall to produce the final output.
    - LLMs have finite model capacity and must prioritise where they allocate their attention. If a prompt asks them to perform many actions, they may prioritise some over others to the detriment of the output.
    - Decomposing tasks can mitigate risk by enabling review points to be placed at key points in the workflow, where risk may be greatest.
  - **Use an ensemble of LLMs where appropriate.** LLMs can be combined to mitigate their limitations:
    - **One LLM reviews the work of another.** Different LLM models may have different capabilities and biases, so independent review can identify issues that may evade the original LLM. Even where the same LLM is prompted to review a previous output, it can be prompted with different priorities and objectives than it had when generating the output, enabling meaningful review. This review can then form part of a new prompt to generate an enhanced version of the output or inform a system decision as to whether to accept or reject the output, or escalate it for human review.
    - **Multiple LLMs attempt a task.** If the same LLM is asked to perform a task a number of times, and then an LLM or human reviewer selects the best output or synthesises them, the sensitivity to the probabilistic nature of the LLM is mitigated. Further mitigation of risk is possible if different LLM models are used, to leverage their different capabilities and biases.

Mitigation of risk of deficient output

System design and development

  - **Establish appropriate human in the loop review and oversight.** The inherent limitations of LLMs mean there will always be some risk that their outputs contain deficiencies. Review and/or oversight by humans of AI systems can mitigate the risk that these deficiencies occur, or that they affect the quality of the audit.
  - **Implement other components to monitor the outputs of GenAI components.** Rules-based processing or predictive machine learning components can evaluate the outputs of LLMs to provide additional mitigation against deficiency. For example, a rules-based component may check that an output meets the requirements of a schema.

- **Appropriate GenAI components.** There are many LLMs commercially available, each with different attributes, and many can be fine tuned or configured further in a range of ways. Selecting the right model, and customising it appropriately, for each use case can significantly mitigate risk. We will now discuss these choices in more detail:
  - **Model choice.** LLMs vary along many dimensions, including:
    - **Reasoning capability.** Some models are optimised for multi-step, structured reasoning, which some use cases demand, for example orchestrator LLMs in agentic systems. Time and cost per output are usually higher for these models, so it may not be proportionate to choose them in every situation.
    - **Domain expertise.** Models may have different areas of expertise, as they are not all trained on the same data. For some use cases, audit and accounting expertise, or knowledge of a specific sector or language, may be especially relevant.
    - **Reliability.** Some models are more consistent and stable in their outputs, and are less prone to sounding confident when they are not. This often mitigates risk, but some tasks may benefit from greater creativity and diversity of output.
    - **Model capacity and context window.** Larger models can learn and apply more patterns, of greater complexity, and consider and integrate more information when generating an output.
    - **Tool aptitude.** Some models are more adept at tool use, which may be relevant for some roles or use cases.
    - **Customisability.** Some models can be fine tuned or configured to a much greater extent than others. This can allow models to be tailored to the demands of specific use cases, which can improve performance and mitigate risk, though default models may be

Mitigation of risk of deficient output

System design and development

  - **Efficiency.** Some models have higher time and cost per output. For demanding use cases, these may well be merited but cheaper, lightweight models may perform comparably on many tasks. Some systems will include both, allocating complex tasks to heavier weight models with lighter models supporting.

- **Model fine tuning.** Some models may be fine tuned by the audit firm, augmenting their training with further training on a curated data set. This can impart domain specific expertise, and promote certain behaviours, but is not without its own risks and costs. For example, fine tuning may not be permitted for the latest models; it can worsen model performance on areas outside the scope of the fine tuning; it can undermine any reliance placed on the often extensive testing performed by the model vendor and the cost to collate the training information and train the model may be significant. In many cases, granting access to tools or prompt engineering may be a preferable way of enhancing domain specific performance or promoting behaviours.
- **Model configuration.** Models may have various configuration options that affect how the model behaves, including token budget; reasoning mode and temperature, which affects the determinism of the model. Configuring the model appropriately can mitigate risk. For example, if an LLM is prompted to summarise a large amount of complex information, it is important that it has appropriate tokens with which to express its output, or it may oversimplify or omit key information.

- **Tool use.** Tools can significantly mitigate many GenAI component performance risks.
  - **Retrieval augmented generation components.** Retrieval augmented generation is a technique whereby information is retrieved from a source and placed in the context window, meaning it can inform the output. Sources can be external, such as auditing or accounting standards, or internal resources. With this technique, LLMs can gain domain specific expertise in areas where their training data may be sparse, in a modular, targeted manner. This partially alleviates the dependence of LLMs on their training data.

Mitigation of risk of deficient output

System design and development

- **Internet search engines.** Using these tools, LLMs can formulate search queries to an internet search engine, with results usually returned as an output or placed within the context window to inform the generation of an output by the LLM. The latter would technically be a form of retrieval augmented generation. These tools let AI systems dynamically access information beyond their training data to ground their responses, though information quality risk may be higher than for information from internally curated sources.
- **Computational tools.** LLMs are often not natively strong at computational tasks, as they lack semantic awareness. If they have access to these tools, however, LLMs can write code, or another structured output, to request the tool to compute something, for example the solution to a mathematical equation, and then receive the output. This significantly mitigates risk compared to the LLM attempting the computation itself.

- **Prompt engineering.** Prompt quality significantly affects GenAI component performance risk; these prompts may be written by humans in the loop; selected from a prompt library or passed to the LLM by the system, specifically by a prompt management component or workflow engine. The quality of human in the loop prompts may be enhanced by the construction of a curated prompt library.

  System prompts are either written by the designers or programmatically constructed from templates and contextual information, such as the goal of the system, what the current step is and relevant outputs from other steps or components. The attributes of a good prompt, whether written by a human or constructed by the system, are discussed later, though for system prompts in particular it may be relevant to note that the structure of output is especially important when the output of an LLM may form part of the input for another.

Mitigations for other component performance risk

The design and development mitigations for other component performance risk are largely the same as for a technological system that does not involve generative or agentic AI. Therefore, we have not included them in this guidance, though many of the mitigations mentioned above as cross-cutting are relevant.

Mitigation of risk of deficient output

System design and development

Mitigations for combination risk

- **Workflow design.** One important element of workflow design is building in review activities at the right points, either by humans, LLM components or rules-based processing components. If errors are not identified early, they can multiply and transform and may mean entire workflows have to be repeated; timely review may significantly reduce the amount of steps that must be repeated. Another mitigation is to prune any unnecessary steps, as each may be an opportunity for amplification or distortion.

- **Rules-based protocols.** In the section on cross-cutting mitigations, we mentioned that defined schemas for component inputs and outputs, rules that govern iteration and response to error can mitigate a range of risks. One category of risk they are particularly effective at mitigating is combination risk. Some other rules-based protocols are primarily combination risk mitigations, including protocols that ensure components communicate effectively with each other. One example is Model Context Protocol, which standardises how AI systems communicate with external systems about inputs, outputs and capabilities, mitigating the risk that information is lost.

  Another form of rules-based protocol that can mitigate combination risk is a rule that the system must maintain a record of the current state of the system, the goal and key intermediary outputs. This mitigates the risk that components stray too far from the context that they should be working within.

Mitigations for information quality risk

The main mitigation for this risk is to provide access to complete and accurate information sources, and catalogue them to support consistent retrieval.

Mitigation of risk of deficient output

Certification

ISQM (UK) 1 requires that firms design, implement and operate systems of quality management that provide reasonable assurance that the firm and its personnel conduct engagements in accordance with professional standards.[1]

Firms often therefore implement processes to certify that technological tools that are to be used on engagements, including AI tools, are working as intended. Certification processes often comprise evaluation of the inputs to the tool, evaluation of the system architecture and component logic of the tool and testing of the outputs.

AI tools are often extremely versatile, with numerous potential use cases. This guidance presumes the unit of certification is an AI tool in the context of a use case. It is a matter of professional judgement how the firm defines each use case; broad definitions may mean fewer certification processes are required, but each may involve more work to obtain the same level of assurance. If a tool, or a range of similar tools, are to be used on similar use cases, it may be possible that some certification activities form part of multiple certification processes.

Generative and agentic AI tools may range from a single LLM to an agentic system with multiple components; as mentioned, certification is for a tool in the context of a use case, and it is a matter of professional judgement to what extent individual components of multi-component systems are evaluated and tested.

System design and development and certification activities, in particular testing, may inform each other iteratively, rather than being performed sequentially. Certification can mitigate a range of risks, including risk of non-compliant methodology, but this section focuses on how it can mitigate risk of deficient output.

Evaluation of system inputs

System inputs usually fall into one of three categories: LLM training data, prompts and information sources accessed by the system at runtime.

- **LLM training data.** It is usually not possible for the audit firm to obtain, and therefore evaluate, this data as LLMs are usually obtained from third party providers, who may not share this information. This does not inherently preclude the use of LLMs, but may have implications for the nature and extent of testing performed in the certification process. If the firm has performed any fine tuning, the certification process may evaluate the appropriateness of that data for training the LLM in the intended manner.

1 ISQM (UK) 1, 14a

Mitigation of risk of deficient output

Certification

- **Prompts.** These may be written by humans in the loop at runtime, selected from a prompt library or passed to the LLM by the system. Human in the loop prompts cannot be evaluated as part of the certification process. Prompts from a library and system prompts may be evaluated as to their clarity, precision and alignment to their intended purpose, though system prompts may be more appropriately evaluated as part of the system architecture and component logic.

- **Information sources.** These are information sources accessed by the system at runtime. The certification process may evaluate whether they are complete, accurate, relevant and, where applicable, appropriately catalogued. If this is not performed, or not possible, this may have implications for the appropriate nature and extent of any review of the outputs.

Evaluation of system architecture and component logic

- **System architecture.** The certification process may evaluate whether the system is appropriately constructed to consistently produce outputs of appropriate quality. This may involve evaluating the system against the risk mitigation ideas discussed in the system design and development section.

- **Component logic.** Many of the components in generative or agentic AI tools may be obtained from third parties. This may mean detailed information about the internal logic of those components is not available for the audit firm to evaluate, though it may be possible to obtain third party assurance that it is sound. Where possible, for example for components that the firm has built itself, the certification process may involve evaluation of whether the internal logic of components is appropriate for their intended purpose.

  If the firm cannot evaluate the internal logic of a component, or obtain assurance that it is appropriate, this does not inherently preclude the use of the component, but may have implications for the nature and extent of testing the firm may choose to perform in relation to that component.

Mitigation of risk of deficient output

Certification

Testing the outputs

Testing that the tool consistently produces outputs that are appropriate for the intended purpose is an important part of the certification process. This may carry greater weight than in some other tool certification processes due to the potential limitations around evaluating the inputs to LLMs, and their internal logic. The certification process may include testing of tool performance in:

- Default expected use cases;
- Edge cases;
- Diverse cases, as the opacity of generative and agentic AI tools can mean it is challenging to predict where they may perform less well;
- Cases that test known potential issues, for example overconfidence or known biases or expertise gaps of the LLMs in the system.

Tests will often be qualitative in nature, especially in relation to the outputs of LLMs. These may be enhanced by the application of structured criteria in evaluating the outputs, to mitigate against bias or inconsistencies.

The tolerance for deficiencies in system outputs is a matter of professional judgement, both in terms of their quantity and individual significance, not least because they will often be qualitative in nature. There is an interrelationship between tolerance for deficiencies, the nature and extent of the human review and oversight that will be performed and how the outputs are to be used on audits, noting that audits provide reasonable, not absolute, assurance. Where tools comprise multiple components, the certification process may test individual components as well as the tool to understand their behaviour and identify potential weak links in the system. This may inform:

- Revisions to the system design;
- Where system tests are targeted;
- The nature and location of human in the loop review and oversight points.

Third party assurance may provide valuable information about the performance of a component or tool but, even if obtained, the firm may perform its own testing on the performance of the component or tool in the context of the relevant use case.

Mitigation of risk of deficient output

Certification

Recertification

The firm may choose to recertify a tool for a range of reasons, including:

- If there are updates to components or system inputs that materially alter the performance of the tool;
- If the tool is to be used on a materially different use case;
- If performance issues are observed;
- If there are relevant revisions to auditing standards, regulatory guidance or internal policies;
- If a certain amount of time has passed.

Monitoring

Implementing a process to monitor the performance of the tool, including the quality and consistency of outputs, patterns of failure and any unexpected behaviour, and how this affects audit quality, may provide valuable information about if, how and when the tool should be used; what further training or guidance may be required; the design of the tool and whether to recertify.

Limited deployment

Firms may choose to release a tool to a limited number of teams to obtain feedback, prior to full certification. There may be greater risk of deficient outputs from these tools as, for example, training, guidance and prompt libraries may not yet have been created or refined. This may have implications for the reliance it may be appropriate to place in these outputs, and teams often perform alternative procedures in full, in parallel to the AI-enabled procedure.

Mitigation of risk of deficient output

Staff education and governance

Educating staff and implementing policies on how and when generative and agentic AI tools should be used is an important mitigation for risks of deficient output, discussed in this section, and misuse of output.

Staff education

The firm may provide staff with training and guidance on how and when to use an AI tool:

- **When to use the tool.** Generative and agentic AI tools are often built for specific use cases, and outside these they may produce deficient outputs or fail to work at all. Others may be relatively versatile, but this does not mean they can be deployed without thought as to whether they are appropriate for the situation. The firm can mitigate the risk of deficient outputs by clearly communicating criteria that should be met for the tool to be used. Information about when the tool may not perform well, and should not be used, may be obtained through the certification process.
- **How to use the tool.** Training and guidance on appropriate system inputs, and how to review and oversee the tool, can mitigate the risk of deficient outputs.
  - **Human in the loop system inputs.** Prompts are the main form of human in the loop system input for generative and agentic AI tools. The quality of prompt, both in terms of task specification and the provision of supporting information, can significantly affect the quality of output.
    - **Task specification.** The prompt should clearly and precisely communicate the task, which should be aligned with the intended purpose of the tool, and be an appropriate amount of work for the tool to perform in one go. The prompt may contain instructions on how to approach the task, for example tools or reasoning modes that it should use, or steps that it should perform.

      The prompt may specify the form of the output, and content it should include. For example, it may tell the tool that, if it is uncertain about parts of the output, it should communicate this.

Mitigation of risk of deficient output

Staff education and governance

- **Supporting information.** The prompt may include or attach supporting information, for example blueprints to follow, context for the request, what the priorities should be, if it faces competing objectives or any other reference material. Organising the supporting information in a structured format may enhance the influence it has on the tool's behaviour.

– **Review and oversight.** Issuing training or guidance for staff on how to review and/or oversee the work of the tool can mitigate risk by improving the quality of outputs and identifying deficiencies so that audit quality is not affected. This may include training and guidance on the purpose of the tool, as this is important context for their judgement as to whether that purpose has been achieved, what oversight actions are appropriate and the risk of automation bias and strategies for avoiding it. Educating staff on known performance issues with the tool may enhance their review and oversight. For example, if the tool is known to sound confident even when it is wrong, it is important that staff are aware of this and apply appropriate professional scepticism.

Staff education is an ongoing activity, rather than something which is only relevant at the launch of a tool. The quantity of material produced is a matter of professional judgement and may vary depending on the use case and the other mitigating activities performed.

Governance

The firm may implement policies on when and how generative and agentic AI tools may be used. These may vary in their prescriptiveness on a case-by-case basis.

- **When to use the tool.** The firm may choose to curate a narrow range of use cases on which staff are permitted to use a tool, or a prescriptive set of criteria that must be met for the tool to be deemed appropriate. Alternatively, it may leave staff to apply their professional judgement to a greater extent in determining when it may be appropriate to use the tool. The former may mean it is easier to design, develop and test the tool, and produce educational material on its use, which may lead to greater mitigation of risk. However, the fewer the use cases, the less potential value may be obtained from the tool compared to if more use cases or flexibility were permitted.

Mitigation of risk of deficient output

Staff education and governance

- **How to use the tool.** The firm may control to a greater or lesser extent how a tool may be used. For example, configuration options may be set centrally or by the audit team; prompts may be written by the designers, required to be from a prompt library or written by the audit team and how a tool should be reviewed or overseen may be specified in methodology or an auditor judgement.

  In relation to configuration and prompts, greater control generally mitigates more risk, at the cost of versatility; for example, prompts from a library may have been written by specialists and validated as part of the certification process, in contrast to those written by the audit team, but even an extensive library cannot match the diversity of bespoke prompts that could be written by audit teams. In relation to review and oversight, it is likely that some requirements will be imposed by the methodology on what this should entail for all use cases, with further specification for some.

The presence of the factors below may lead the firm to be more prescriptive regarding when and/or how a tool is used, but significant professional judgement is required.

- The tool has been built for a specific use case
- There is a significant risk to audit quality if outputs are deficient
- Prompts written by the human in the loop produce materially worse outputs than those written by professional prompt engineers
- Testing has been narrowly focused on a few prompts, configurations and use cases
- Testing shows that performance is sensitive to variations in prompt, configuration or use case
- Pilot deployment shows that audit teams do not consistently use the tool appropriately, or in the appropriate situations
- Output deficiencies may not be reliably identified by review or oversight
- Review and oversight must include certain activities or be performed by those with a certain skillset to be effective.

Mitigation of risk of deficient output

Human in the loop review and oversight

Review

Review of the outputs of generative and agentic AI tools by humans in the loop can mitigate the risk that deficiencies in outputs are not identified and therefore that audit quality is affected. When we refer to review, we mean review of the final outputs of generative or, to the extent applicable, agentic AI tools; review of intermediary outputs of agentic AI tools is discussed in the oversight section below.

To be effective at mitigating risk, reviews should be performed by appropriate staff and comprise appropriate activities.

- **Appropriate reviewers.** It is important that those who review the outputs of generative and agentic AI tools have the appropriate competence to identify potential deficiencies in the outputs. This may require expertise in a range of areas, depending on the use case. For some tools and use cases, it may be that audit and accounting expertise is sufficient, though knowledge of a specific area of the relevant audit engagement may provide important context that may enhance the review. For others, specialist expertise may be required.
- **Appropriate review activities.** The review should evaluate whether the outputs are appropriate for the intended use, complete, accurate, internally consistent and consistent with the reviewer's understanding of the entity. The reviewer should apply professional scepticism and their knowledge of the main risks of output deficiency for this tool and use case, and be aware of the risk of automation bias.

  The review may involve forming an independent view to compare some or all of the tool output against. For example, review of a GenAI summary may involve reading the source material and confirming that key information has been accurately summarised, and review of an AI generated list of risks of material misstatement may involve the auditor exercising their professional judgement to identify potential risks of material misstatement themselves.

  The nature and extent of review is a matter of professional judgement based on residual risks of deficient outputs after other mitigating activities, how these risks may be mitigated by review and the risk to audit quality if a deficient output is not identified.

Mitigation of risk of deficient output

Human in the loop review and oversight

Oversight

Oversight activities in relation to agentic AI tools can mitigate the risk that a tool produces deficient outputs. Oversight activities may be diverse but may include reviewing intermediary outputs of the tool, authorising the system to continue or perform certain actions or choosing what action the tool should perform. Professional scepticism and an understanding of automation bias are important elements of appropriate oversight.

Oversight activities occur at control points, when specified criteria are met. For example, this may be when certain intermediary outputs are ready, the tool wishes to perform certain actions or after a certain amount of time or activity.

The choice of criteria, and therefore where control points are, is an important professional judgement. It may be based on the nature of the main risks of output deficiency, which may be informed by the certification process, how oversight may mitigate these risks and the potential risk to audit quality if output deficiencies are not identified.

Mitigation of risk of misuse of output

The risks that outputs are misinterpreted, or misused due to a misunderstanding of the methodology, can be mitigated through appropriate review by other members of the audit team, and by educating staff on how to interpret the outputs of the tool, including their scope and limitations, and how the methodology expects them to respond depending on the nature of the output. It may be easier to provide this education if there are a relatively focused set of use cases.

Regarding risk of misinterpretation specifically, another mitigation is to build the system so that outputs are explained. For example, the output may include the chain of thought that led to the output, including references to the information considered.

Mitigation of risk of non-compliant methodology

This risk may be mitigated by involving colleagues from methodology teams throughout development. This risk may arise if the methodology misconstrues the nature of the outputs of the tool, or what may be inferred from them.

Collaboration between methodology and technology teams may mitigate this, by facilitating personnel in methodology teams obtaining an enhanced understanding of how the tool works and the nature of the outputs, and in particular their limitations, enabling them to write methodology that appropriately guides audit teams on how to use the outputs to perform audits that meet auditing standards. Further, personnel in technology teams can understand how the tool will be used and relied upon, which may provide important context as they develop the tool.

The certification process for a tool may validate the supporting methodology against auditing standards.

Illustrative examples

Illustrative examples

Introduction

These illustrative examples portray the considerations of a hypothetical firm of the risks to audit quality posed by the use of a generative and/or agentic AI tool in specific use cases, and how they may be mitigated.

These examples show one way that these tools may be designed and implemented, which may not be the best or only approach. They do not include every relevant risk and mitigation that a firm may or should consider or implement in a real deployment.

Illustrative example 1

Summarisation of board minutes

Use case

Auditors often review the board minutes of an audited entity as part of their risk assessment, to enhance their understanding of the entity and identify potential risk indicators. The firm has identified a use case for generative AI to enhance the quality and efficiency of this work by summarising the minutes and identifying areas of the most potential relevance for the auditor. This does not replace the auditor's risk assessment, but rather enhances it by enabling the auditor to focus their time on the most relevant areas and potentially identifying insights the auditor may miss. The main component of the tool is an LLM, with other components including a component that chunks and indexes the minutes, a prompt management component and a user interface. The auditor documents on the audit file that they have used a generative AI-enabled tool to focus their review of board minutes, and how the output informed any further work they performed and their conclusions in respect of the board minutes. The remaining documentation in relation to this tool is retained centrally, in line with the FRC AI in the Audit Guidance.

Illustrative example 1

Summarisation of board minutes

Risks of deficient output

Risk	Mitigation
The tool omits information	**System design and development** The elements of system design and development that are most relevant for mitigating these risks relate to workflow design, choice and configuration of components and prompt engineering. • **Workflow design.** The workflow is designed to mitigate these risks in two main ways. Firstly, it includes a step prior to summarisation where the minutes are segmented and indexed with stable identifiers. This enables the tool to cite its sources consistently and accurately. This reduces the risk of hallucination, as the LLM is less likely to fabricate information if it is required to cite a source for everything it includes in the summary, and supports effective human review of the output, which is a mitigation for all three of these risks. Secondly, the summarisation itself is comprised of three distinct activities, each prompted separately. The minutes from each meeting are first summarised individually based on the prompt; then all of the minutes are reviewed collectively with a single prompt for themes that may only emerge in the context of the minutes for the whole period and finally the outputs of the first two steps are synthesised into a final summary. This distributes the cognitive load of summarising a potentially extensive amount of complex information across multiple steps, mitigating the risk that information is omitted or distorted.
The tool distorts information	
The tool hallucinates information	

Illustrative example 1

Summarisation of board minutes

Risk	Mitigation
The tool omits information The tool distorts information The tool hallucinates information	• **Appropriate components.** LLMs vary along multiple dimensions and, to mitigate the risks identified, the firm chooses a model that has relatively strong domain expertise in audit, accounting, relevant languages and business sectors; is relatively reliable in its outputs and has sufficient context window size to reason over the full set of board minutes. In addition, the firm grants the LLM sufficient token budget to generate its output, to reduce the risk that information is overly simplified or omitted. • **Prompt engineering.** This tool uses prewritten prompts that are fed to the LLM by a prompt management component. The quality of these prompts affects the quality of the outputs, and specifically their susceptibility to omissions, distortions and hallucinations. The prompts provide clear instructions on what matters may be relevant for the auditor, and therefore should be included in the output. The prompts tell the LLM to: – Structure the output around the headings in the minutes; – Err on the side of including matters if it is not sure if they are relevant; – Include something in the summary only if it is a summary of something in the board minutes; – Cite the material in the minutes that each part of the summary is based on, with reference to the stable identifiers assigned prior to summarisation. Further, the prompts include examples of what the output should look like. If the auditor wants to alter a prewritten prompt, they can consult with a central technical team to determine whether this would be appropriate. **Certification** The certification process includes robust testing of the tool. The firm intends to deploy the tool across its UK practice, so tests cover how the tool performs in the context of minutes from entities of different sizes and sectors, and the prompts are refined based on some observed deficiencies in early testing. Perfect accuracy in these tests is not required for the tool to be certified, see the section on appropriate confidence in the quality of outputs for further discussion.

Illustrative example 1

Summarisation of board minutes

Risk	Mitigation
The tool omits information The tool distorts information The tool hallucinates information	**Staff education and governance** The training and guidance focuses on how to review the outputs of the tool, as there are no judgemental criteria for when to use the tool and the auditor does not prompt the tool themselves. The training and guidance reminds staff of the importance of professional scepticism, the risk of automation bias and the tendency of the tool to sound fluent and confident even when it is wrong, and cautions staff about the risks of output deficiencies observed in testing. The firm exerts significant control over how the tool is used, in that the auditor must use prewritten prompts from a library. This mitigates the risk of deficient outputs as the prompts are written by a specialist team and tested as part of the certification process. **Human in the loop review** The firm methodology requires the auditor to review the output of the tool and corroborate its completeness and accuracy by reading the board minutes themselves. A more detailed review of the cited board minute sections is required in relation to areas of the output that are not consistent with the auditor's prior understanding of the entity, or have implications for their risk assessment. The firm guidance suggests that the review is performed by a member of the engagement team that has an appropriate understanding of the entity.

Illustrative example 1

Summarisation of board minutes

Appropriate confidence in the quality of outputs

The confidence it is appropriate to obtain in the quality of an output of a tool is sensitive to how it will be used and relied upon. In this case, the output is used to facilitate a more focused review of the board minutes and potentially the identification of insights that the auditor may have missed, and therefore an enhanced understanding of the entity and risk assessment.

The board minute review is not the only procedure the auditor will perform to understand the entity or assess risk. However, relying on an output of this tool that contains deficiencies may result in a risk assessment that is not appropriate and the confidence that the firm deems it appropriate to obtain is set accordingly.

Significant professional judgement is exercised in determining how mitigating activities should be combined to obtain appropriate confidence.

The design of the tool is relatively simple, in contrast to an agentic system with multiple LLMs, for example. It may have been possible that a more complex system would produce higher quality outputs and, theoretically, that these may have been of consistently high enough quality prior to human in the loop review that a scaled back review, for example one that did not involve the human reading the board minutes, may be sufficient to obtain appropriate confidence. It was determined, however, that there was significant uncertainty over whether this consistency of quality could be achieved with current models, and that the approach chosen would be a more efficient means of obtaining appropriate confidence.

The nature and frequency of output deficiencies observed in initial testing informs the nature and extent of human in the loop review that the methodology will require. This, in turn, informs the tolerance for output deficiencies for the tool to be certified. The tolerance would have been lower, for example, if the human in the loop was not required to read the board minutes as part of the review.

The firm opts for a combination of mitigations, including those discussed in this example, that it judges enable an auditor using the tool as intended to have appropriate confidence in the quality of its outputs.

Illustrative example 1

Summarisation of board minutes

Risks of misuse of output

The main risk of misuse of output that the firm identifies is that auditors may interpret the output as a complete summary of the entity's activities in the period. This may lead to them not performing appropriate other risk assessment procedures.

The mitigation the firm implements for this risk is to ensure that the training and guidance that auditors receive in relation to this tool clearly explains the scope of the output, and how it fits into a full risk assessment.

Risks of non-compliant methodology

The firm assesses this risk as relatively low as the AI enabled review of board minutes is simply an enhancement of a procedure that auditors were already required by the firm's methodology to perform. The firm's methodology with respect to the rest of the risk assessment is not revised, so there is no reduction in expectations around the quality of other risk assessment procedures.

Illustrative example 2

Contract review

Use case

When auditing revenue, auditors often test a sample of contracts to obtain sufficient appropriate audit evidence that revenue has been recognised in accordance with the applicable accounting framework. The sample selection is often risk based, and the firm has identified that AI may be able to improve the quality of this procedure by enhancing the targeting of the sample selection on contracts in relation to which there is a greater risk of revenue being misstated due to revenue recognition issues. In the remainder of this example, we refer to these as higher risk contracts.

The tool is comprised of a workflow engine, an LLM, a component that chunks and indexes the contracts, storage, a retrieval augmented generation retrieval module, a computation component and a user interface. The workflow of the tool is as follows. Uploaded contracts are first chunked and indexed. Then, the LLM, under instruction from a prewritten prompt, reviews each contract and extracts certain specified content that may indicate risk.

Some of the extracted pieces of content will themselves be risk indicators that, if present, are unambiguous, for example the presence of terms such as bonus, contingent payment or refund or a contract length over a certain limit. The prompt will tell the LLM to label these as the relevant indicator in its output for that step, and the tool then records their presence or absence in each contract for a later step.

Other pieces of extracted content will require evaluation to determine the extent to which they indicate risk, for example whether performance obligations are distinct, whether control transfers over time or whether variable consideration is constrained. The extraction prompt tells the LLM to label these as relating to the relevant indicator. Then, the material relating to each of these indicators is collated and evaluated, with a prewritten, bespoke prompt for each indicator.

The outputs of these evaluations for each indicator and contract are recorded with the unambiguous risk indicators, and then the computation component calculates a risk score for each contract. Contracts with a score above a defined threshold comprise the sample that the auditor tests. The auditor can see which risk indicators were present in each contract.

Illustrative example 2

Contract review

The tool exhibits some agentic behaviour; it can perform multiple steps independently and determine, for example, whether to escalate an issue to human review. However, the agency is bounded within an entirely human-designed workflow.

The auditor documents on the audit file that they have used an AI-enabled tool to identify a sample of higher risk contracts, their assessment against the criteria for use of the tool to be appropriate, what contracts were selected, the risk indicators these contracts exhibited and the substantive testing performed. The remaining documentation in relation to this tool is retained centrally, in line with the FRC AI in the Audit Guidance.

Illustrative example 2

Contract review

Risks of deficient output

1. Extraction – The output is deficient because the prompt's specification of what content the LLM should extract is not fit for purpose
2. Extraction – The output is deficient because the LLM fails to extract all and only the content specified in the prompt
3. Evaluation – The output is deficient because the LLM misevaluates the extracted content
4. Calculation – The output is deficient because the risk indicator weighting or score threshold is not appropriate

Mitigation category	Mitigation	Risks mitigated
System design and development – workflow design	The workflow includes a step prior to extraction where the contracts are segmented and indexed with stable identifiers, enabling the tool to cite content consistently and accurately. This reduces the risk of hallucination, as the LLM is less likely to fabricate information if it is required to cite sources for the content it extracts, and supports effective human review of the output by providing the reviewer with citations for what led to the contract being identified as higher risk.	2, 3
System design and development – workflow design	The cognitive load of extracting information and then evaluating it is distributed across separate steps, with separate prompts. This mitigates the risk that output deficiencies result from compute or model capacity limitations. Further, it enables escalation to a human in the loop at both the extraction and evaluation steps, if appropriate.	2, 3
System design and development – workflow design	Splitting extraction and evaluation allows the LLM to be prompted to behave in the appropriate manner for each activity, enhancing performance on both.	2, 3

Illustrative example 2

Contract review

Mitigation category	Mitigation	Risks mitigated
System design and development – appropriate components	The firm chooses a model that has relatively strong domain expertise in audit, accounting, relevant languages and business sectors; is relatively reliable in its outputs; has sufficient context window size to reason over the entire contracts and is adept at tool use. In addition, the firm grants the LLM sufficient token budget to generate its output, specifically when extracting content, to reduce the risk that information is simplified or omitted.	2, 3
System design and development – prompt engineering	The prompt writing team consults with relevant specialist technical audit and accounting teams so that: • The extraction prompt includes the appropriate categories of content that may indicate risk; • The extraction prompt reflects that contracts may differ in their use of terminology; • The evaluation prompts include appropriate criteria for evaluating the extent to which the content indicates risk.	1, 2, 3
System design and development – prompt engineering	The extraction and evaluation prompts are clear and precise with respect to what should be extracted and how it should be evaluated. They tell the tool to request human judgement if it is not confident on how to apply the extraction or evaluation criteria to a specific piece of content. The prompts tell the LLM to cite where material has been extracted from, with reference to the stable identifiers assigned prior to extraction, and which extracts were most relevant in arriving at the evaluation. Further, the prompts include output schema and examples of what the outputs should look like.	2, 3
System design and development – tool use	Evaluating the risk indicated by extracted content can be highly judgemental and technical. The firm curates and indexes relevant intellectual resources, that can be retrieved through a retrieval augmented generation component, to enhance the domain expertise of the tool and support consistent and appropriate evaluation.	3

Illustrative example 2

Contract review

Mitigation category	Mitigation	Risks mitigated
System design and development – mitigations for other component performance risk	The design and development team consults with relevant technical audit and accounting teams on the appropriate weightings for each risk indicator and score threshold for a contract to be deemed higher risk. These teams have the technical expertise to understand, and empirical experience of, how reliably each indicator indicates risk. This mitigates the risk that weightings for each indicator are not representative of the extent to which they indicate risk, or that the threshold is not appropriate for obtaining reasonable assurance in the context of the relevant technical team's professional judgement of how representative of risk the indicators are, the performance of the tool at extracting and evaluating content, the substantive testing is performed in relation to higher risk contracts and any other procedures performed.	4
Certification – testing the outputs	The performance of the LLM component with each prompt is individually tested. This evaluates the appropriateness of the categories of content that are included in the extraction prompt, the LLM's accuracy at extracting content and its ability to evaluate the level of risk indicated. Certification includes testing in the context of contracts from different sectors, contracts that differ in their use of terminology, contracts that exhibit each of the risk indicators and contracts with high and low numbers of risk indicators. Some intentionally ambiguous cases are tested, to see if the LLM would escalate them to human review.	1, 2, 3
Certification – evaluation of system inputs	The risk indicator weightings and score threshold are tested theoretically and empirically. The certification team calculates whether hypothetical and real combinations of risk indicators would combine for a score that exceeds the threshold and evaluates, in collaboration with the relevant technical teams, whether this appears appropriate in the context of how the output is used.	4

Illustrative example 2

Contract review

Mitigation category	Mitigation	Risks mitigated
Certification – testing the outputs	The tool as a whole is tested in the context of a diverse and representative range of contracts. For discussion on how effective the tool must be to be certified, see later section on appropriate confidence in the quality of outputs.	1, 2, 3, 4
Staff education and governance – staff education – when to use the tool	The firm provides training and guidance to teams on the criteria relating to the engagement, entity, revenue and contracts that must be met for it to be appropriate to use the tool.	1, 2, 3
Staff education and governance – staff education – how to use the tool	The firm provides training and guidance on the substantive testing the firm methodology requires the auditor to perform over the higher risk contracts, and tool functionality such as the ability to see why a contract has been identified as higher risk. The substantive testing performed in relation to a contract are partially sensitive to the risk indicators that led to that contract being identified as higher risk. The training and guidance communicates to teams that, when performing substantive testing, they should remain alert for anything that may indicate a systemic issue with the tool's stratification of contracts by risk on that engagement or that there may be greater than expected risk in relation to contracts not identified as higher risk.	1, 2, 3
Staff education and governance – staff education – how to use the tool	The tool can escalate a judgement about extraction or evaluation to the human in the loop. The training and guidance includes criteria to support the auditor in their judgements as to whether something should be extracted, or how it should be evaluated.	1, 2, 3

Illustrative example 2

Contract review

Mitigation category	Mitigation	Risks mitigated
Staff education and governance – governance – how to use the tool	The firm exerts significant control over how the tool is used, in that the auditor must use prewritten prompts from a library. This mitigates the risk of deficient outputs as the prompts are written by a specialist team and tested as part of the certification process.	1, 2, 3
Human in the loop review and oversight – review	The firm methodology does not require the auditor to review the appropriateness of the classification of contracts as higher risk or not on a contract-by-contract basis, but to review whether the output is consistent with their understanding of the entity.	1, 2, 3, 4
Human in the loop review and oversight – oversight	Oversight points are dynamic, in that they occur if and when the tool escalates something to the human in the loop. The LLM is prompted to escalate if it deems that it lacks appropriate certainty as to whether something should be extracted, or how it should be evaluated. Escalating at these points mitigates the risk of deficient outputs in situations that may be more judgemental or where the LLM is aware that it lacks some relevant domain expertise or information.	2, 3

Illustrative example 2

Contract review

Illustrative example 2

Contract review

Appropriate confidence in the quality of outputs

The confidence it is appropriate to obtain in the quality of an output of a tool is sensitive to how it will be used and relied upon. In this case, the output is a sample of higher risk contracts that will be substantively tested to obtain evidence that revenue has been recognised in accordance with the applicable reporting framework. Other procedures may be performed in relation to the same assertions, in addition to this AI-enabled procedure, with the aim of obtaining reasonable assurance that there are no material misstatements.

The firm is aiming for **reasonable assurance,** and knows what **assurance will be obtained from other procedures in relation to revenue recognition**, and can therefore determine the assurance required from this AI-enabled procedure.

The **assurance that the firm deems is obtained from this AI-enabled procedure** is a function of its **confidence that the sample selection effectively identifies higher risk items**, and the **nature and extent of the substantive testing performed in relation to the contracts in the sample.**

The firm knows what **substantive testing its methodology will require in relation to contracts selected by the AI tool** and **the assurance it wants to obtain by performing the AI-enabled procedure.** The firm has extensive experience in the application of professional judgement to determine the amount of assurance obtained from a procedure, based on how effective the sample selection process is at identifying risk and the nature and extent of substantive testing in relation to sample items.

The firm therefore has a strong initial understanding of how effective the tool must be at identifying higher risk items. To enhance the precision of its judgement, the firm benchmarks against other procedures to focus its professional judgement as to how **confident it must be that the AI tool effectively identifies higher risk items** in order to obtain the **assurance it wants**, in the context of the **substantive testing it will require in relation to the contracts selected**.

This confidence, which is the same as the **confidence in the quality of the output of the tool**, is a function of the **confidence in the technical judgements around risk indicator selection and weighting**, **confidence in the performance of the tool at extracting and evaluating content** and the **score threshold for a contract to be selected.**

Illustrative example 2

Contract review

Appropriate confidence in the quality of outputs

Confidence that the risk indicators reliably indicate risk, and are weighted proportionally, is obtained through consultation with the relevant technical audit and accounting teams. These teams have the technical expertise to understand, and empirical experience of, how reliably different risk indicators indicate risk.

Confidence that the tool can consistently extract and evaluate content is obtained through a combination of system design and development, certification, staff education and governance and human in the loop oversight. Notwithstanding the fact that the audit team reviews the consistency of the output with its understanding and stays alert for signs of any systemic performance issues when substantively testing, and that the tool may escalate issues to a human auditor, by default there is no contract-by-contract review of the output by a human. Therefore, the performance of the tool at extracting and evaluating content that is observed in testing is a good proxy for the performance of the tool at these activities in full deployment.

The score threshold for a contract to be deemed higher risk can be calibrated in the context of the confidence in the technical judgements around risk indicator selection and weighting and the confidence in the performance of the tool at extracting and evaluating content. The lower the threshold, all else being equal, the more contracts will be selected and the greater the effectiveness of the tool at identifying higher risk items.

The firm compares the effectiveness of the risk indicators at identifying higher risk items to other risk-based sample selection processes and adjusts for the observed level of performance at extraction and evaluation, which will not be perfect due to the inherent limitations of LLMs. The firm then sets a threshold that it judges will result in a level of confidence that the tool effectively identifies higher risk contracts that it determined will lead to it obtaining the assurance it wants from the AI-enabled procedure.

Illustrative example 2

Contract review

Risks of misuse of output

The firm identifies three main risks of misuse of output in relation to this tool.

The first is that the auditor misinterprets the output and performs substantive testing that is not related to the relevant risks for that contract. This is mitigated by building the tool so that the auditor can see which risk indicators led to that contract being identified as higher risk.

The second is that the auditor may not understand what substantive tests they should perform based on the risk indicators that were identified, and the third is that they may believe that they have no responsibilities in relation to contracts not identified as higher risk, when in fact they must remain alert when substantively testing the higher risk contracts for signs that there may be greater than expected risk in relation to contracts not identified as higher risk.

Both are mitigated through the training and guidance that auditors receive in relation to this tool.

Risks of non-compliant methodology

The firm identifies the main risk to the methodology not being compliant as the risk that the firm evaluates the tool as more effective than it is at identifying higher risk contracts, and therefore that it overestimates the assurance obtained from the AI-enabled procedure.

This is mitigated through the certification process and extensive collaboration with audit and accounting teams.

Financial Reporting Council

Financial Reporting Council
London office:
13th Floor,
1 Harbour Exchange Square, London,
E14 9GE

Birmingham office:
5th Floor,
3 Arena Central,
Bridge Street,
Birmingham, B1 2AX

+44 (0)20 7492 2300
www.frc.org.uk

Follow us on
LinkedIn