Get on board or get left behind

Visionary Internal Audit practices are charging ahead with advanced generative AI solutions.

Many, if not most, internal audit (IA) teams today have access to generative artificial intelligence (GenAI) tools in their enterprise environments that are approved for routine use. Yet, visionary Chief Audit Executives (CAEs) have blazed past standard out-of-the-box solutions. These leaders understand the revolutionary nature of this technology and are seeking transformational outcomes that have impact beyond incremental personal productivity enhancements. This white paper will explain the techniques these leading professionals are employing to achieve their audacious goals.

GenAI is just one part of a technology-enablement strategy.

GenAI has been riding the hype cycle to ever-increasing new heights. In the echo chamber of GenAI hype, it is tempting to view this powerful technology as an “everything tool” that solves all problems. Innovative auditors view GenAI as one tool on a work bench. It sits alongside analytics, automation workflow tools, audit management solutions, and other technologies that are used in concert to address IA’s goals. While this article focuses on GenAI, it is critical to develop a holistic view of the roles that all these technologies will play in your IA strategy.

Selecting the right objective is essential.

The first, and perhaps most essential, element of any technology-enablement program is to have a clearly articulated, target outcome. The deployment of GenAI is no exception. It is common for IA functions to use metrics as goals, such as of “using analytics on x% of audits,” or to have impossibly vague goals, such as “more insightful audit reports.” These approaches typically lead to unimpressive results and frustrated team members. Audit executives should understand that targeted outcomes must be prioritized and that trade-offs must be made. Analytics may provide a more complete picture of business risks but require more time and cost to deploy. Automation may reduce completion times for tasks but can be costly to develop and maintain. Each IA function will have to evaluate their organizational context to determine how to position themselves within the triple constraint of time, scope, and cost. **Define your target outcomes clearly—ambiguity leads to mediocrity. Visionary CAEs make strategic trade-offs for maximum impact.**

Prompt engineering is a starting point.

Making GenAI responses more relevant to IA generally starts with prompt engineering, which is the practice of creating tailored prompts that will improve the quality of model responses. Most users learn the basics of prompt engineering early in their experience with GenAI applications. However, prompt engineering can become very sophisticated and may require additional training and education to master. Some common prompt engineering approaches are shown here.

Verbose prompting provides more detailed context to the prompt with the goal of making the outputs more useful. Some common elements and examples include:

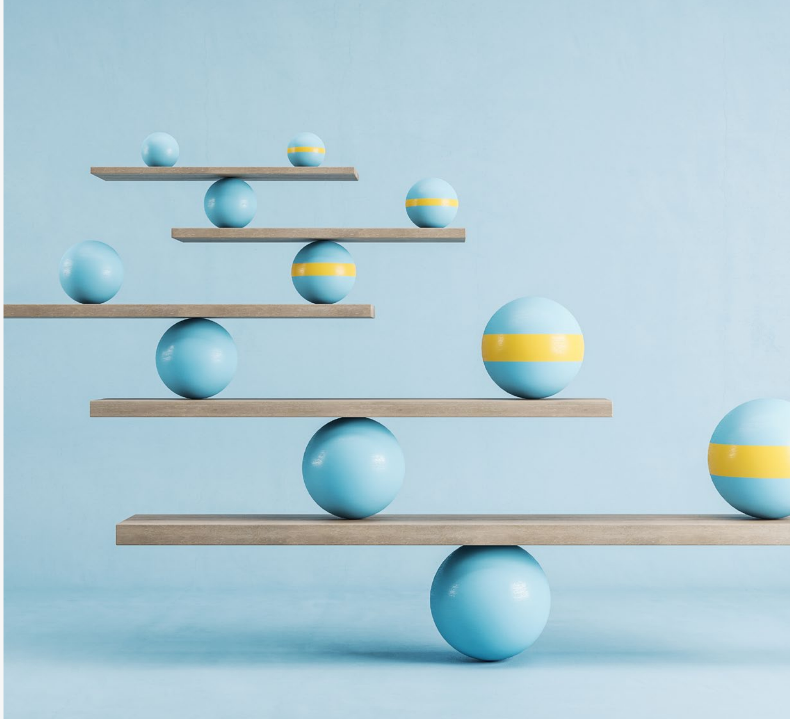
- Role: "As an internal auditor for a large manufacturer…"
- Context: "… I am preparing for an audit of the company's payroll process…"
- Task: "…draft an email to a control owner requesting the following documents…"
- Additional information: "…the control owner should provide the documents within ten business days to audit@company.com...
- Tone: "The email should be written in concise, professional English."
- Constraints: "Limit the email to 500 words."

Few-shot learning is an approach that includes giving examples as a part of the prompt. This allows the GenAI application to match the pattern of its response to the examples provided by the user. This can be used along with verbose prompting to further improve the results. For instance, a user might expand on the example above to include some representative document request emails from previous audits. Depending on the capabilities of your GenAI model the examples may be included as text in the prompt or as files that are uploaded as part of the prompt.

Persona-based prompting is an approach where IA teams will create prompts that ask GenAI applications to take on specialized roles to provide specific perspectives. A prompt is developed that describes a persona, what questions they are focused on, what industry frameworks they may consider, etc. Auditors can start a GenAI conversation by establishing the persona via prompt and proceed to asking that persona to provide insights such as proposing revisions to documents. Combining these different perspectives can improve the quality and relevance of the outputs to stakeholders. Some examples include:

- "CEO" is focused on how this will impact the organization and its strategy.
- "CFO" wants to know about the budgetary impact and compliance with US GAAP.
- "CISO" considers impact in the context of the NIST cyber security framework.
- "Editor" suggests improvements to the writing to improve grammar, tone, and structure.

Other techniques are available to accomplish specific tasks. 'Chain of Thought' prompting seeks to improve outputs that require the model to apply human-like reasoning by adding intermediate reasoning steps into prompts. 'Tree of Thoughts' prompts increase the quality of reasoning tasks by asking the model to evaluate multiple possible responses. 'Chain of Density' prompts can summarize large volumes of information into easily readable texts. These and many other techniques can be useful to IA, and we encourage practitioners to research and experiment with advanced prompting techniques.



Prompt libraries allow everyday users to harness powerful prompts.

With so many techniques available, leading IA functions have begun creating prompt libraries. These libraries are a repository for prompts that all IA team members can use. This approach provides many advantages. It reduces the amount of time spent by individual auditors on engineering prompts for tasks they perform frequently. They can serve as a collection of "best practices" to achieve improved, more consistent results. After all, a big hurdle to effective use of GenAI is lack of imagination by a user. Prompt libraries will give auditors access to complex prompting approaches that they might otherwise not be familiar with or able to use without support. It can enable the transition from an execution mindset to a transformation mindset.

Prompts to be added to the prompt library are typically harvested from IA team members that have been successful using GenAI to perform an audit task. These prompts are often refined with the help of an experienced prompt engineer before including it in the library. Prompt libraries often provide some degree of documentation that explains the function of the prompts and uses key words, audit phases, and other cross-references make it easier to find as the library gets larger. Prompt libraries at various IA functions take different forms from simple spreadsheets to interactive dashboards or applications. In at least one instance, an IA team made their prompt library resemble recipe cards.

Advanced prompt engineering can streamline the user experience.

As prompts become increasingly sophisticated, innovative IA teams have begun to get away from having users type (or paste) prompts. For tasks that auditors frequently perform, IA teams can build an applet that serves as a front-end to the GenAI application. In such instances, the user will fill out a form which may include checkboxes, dropdown menus, free text fields and other controls. The information entered into this form along with any attachments the user may include is then used “behind the scenes” by the applet to produce a prompt using a pre-defined prompt template that is automatically passed to the GenAI application. Such prompt templates can contain incremental information to improve performance, such as a style guide, methodological information, or guardrails to be used when creating the prompt.

Enabling auditors with applets reduces the need for training and embeds strong prompt engineering into a smooth and consistent experience. Creating applets can be done in most organizations with common technologies that are already available. Some development and integration capabilities are required, but such development is generally not a high-cost endeavor.

Some examples of applets can perform detailed document reviews for style and tone, extract information from large documents or collections of documents, synthesize and formalize meeting notes, and perform foreign language translations. There are also business examples of voice-to-text interfaces that allow the auditor to vocally speak answers to pre-defined questions. These spoken inputs to the model are converted to text, then automatically incorporated as a prompt the GenAI model can use to produce a response. This capability will become increasingly common as tools become more conversational and interactive.

Embeddings improve factual output and add proprietary information.

Prompt engineering is critical to improving the relevance of GenAI responses, but this discipline alone has significant limitations for IA teams. Chief among them in the fact that GenAI applications typically do not have access to information inside the organization. The names of specific people, places, and products; documentation of your processes; and access to your data will never be accomplished through prompt engineering alone. In order to get access to specific information about your enterprise, you must turn to embeddings.

Embeddings help to identify information about your enterprise that is relevant to your prompt. Essentially, they serve as a knowledge repository for a GenAI application. This knowledge repository may be a single document, such as a methodology, system document, or regulation. It could be a group of documents such as purchasing contracts or historical IA reports. It could be a database or even a collection of images. Regardless of the type of data in the knowledge repository embeddings work in a similar fashion. The language model in the GenAI application generates the language and reasoning to be used in the answer with factual information drawn from the knowledge repository. The most common technique used with embeddings is called Retrieval Augmented Generation (RAG).

RAG Solutions pave the way to more automated auditing.

RAG solutions vastly improve the usefulness of GenAI to IA teams. The “retrieval” part of RAG typically leverages embeddings to find information that is relevant to the user’s prompt. That information is then passed along to GenAI to better answer the user’s request. The fact that GenAI can ingest enormous amounts of text information and to parse that information semantically creates the ability to execute several tasks commonly performed by auditors, especially the “stare and compare” tasks associated with many internal control tests. So, questions like “does this number match that number” or “did activity X happen before activity Y” can be answered at scale using RAG solutions.

The rise of Knowledge Assistants

RAG solutions also enable the creation of knowledge assistants, chat bots that have access to domain-specific data and information. For example, innovative IA teams have created knowledge assistants to facilitate access to their IA methodology. In this case, when a staff auditor has questions about sample sizes or testing approaches, they can interact with the chat bot rather than search for the information on their own or wait for an answer from a human. Not every example is so simple. Data analytics assistants, regulatory assistants, and business process assistants are all capabilities that are currently deployed or in development by visionary IA teams and vendors.

While the underlying technology that enables knowledge assistants is impressive, the level of effort for end users to create knowledge assistants is minimal. GenAI applications can be configured to use any combination of documents in a set of folders on a computer or collaboration site. At KPMG LLP, for example, our teams can put all of the documents from an internal audit into a collaboration site, then interact with those documents via a conversational GenAI interface. This allows any team member to create a knowledge assistant on virtually any topic for which information is available.

Combining prompt engineering with embeddings yields impactful results.

Some of the most powerful applications to emerge from leading IA practices over the past year involve combining prompt engineering with RAG capabilities. One of the most popular examples that has been developed by multiple leading IA teams is a findings writer. The essential idea is that verbose prompts reference work papers, finding language, and methodology documents to produce factual findings language that is consistent with the IA teams preferred style. While a findings writer is a first step, these innovative IA teams have aspirations to progress to full report writing capabilities. This is a clear leap forward, and there are no major barriers to achieving this objective with current technology.

Parameter adjustment improves responses in more technical domains.

Large language models (LLMs), the underlying technology at the heart of GenAI applications, use a complex set of weighted probabilities to compute which words should be used in which order. These weights can be adjusted for most LLMs. There are multiple approaches to adjust the weights, but the most common approach is referred to as **fine tuning.** The impact of adjusting the weights of these models is that the usage of language changes. It is a classic LLM fine tuning exercise in academia to adjust the parameters of the model to speak in Shakespearean language. Such a change does not significantly impact the facts included in the response, merely the language used to express those facts.

In some industries and domains, technical language is important for clarity and credibility. A GenAI solution that uses everyday language to describe technical phenomena will require significant editing. For example, in the IA profession words like *'significant deficiency'* and *'material weakness'* carry specific meanings that English language synonyms will not convey.

Currently fine tuning LLMs is not being pursued by many IA teams. This largely because advanced prompt engineering and embeddings are lower cost and higher impact in the short-term. It is worth noting however that newer approaches to fine tuning, such as retrieval augmented fine tuning (RAFT) may decrease the cost and expertise required for such solutions in the future. We expect that the most relevant and impactful GenAI solutions in the future will rely on a combination of advanced prompt engineering, embeddings, and parameter adjustment to perform internal audit tasks.

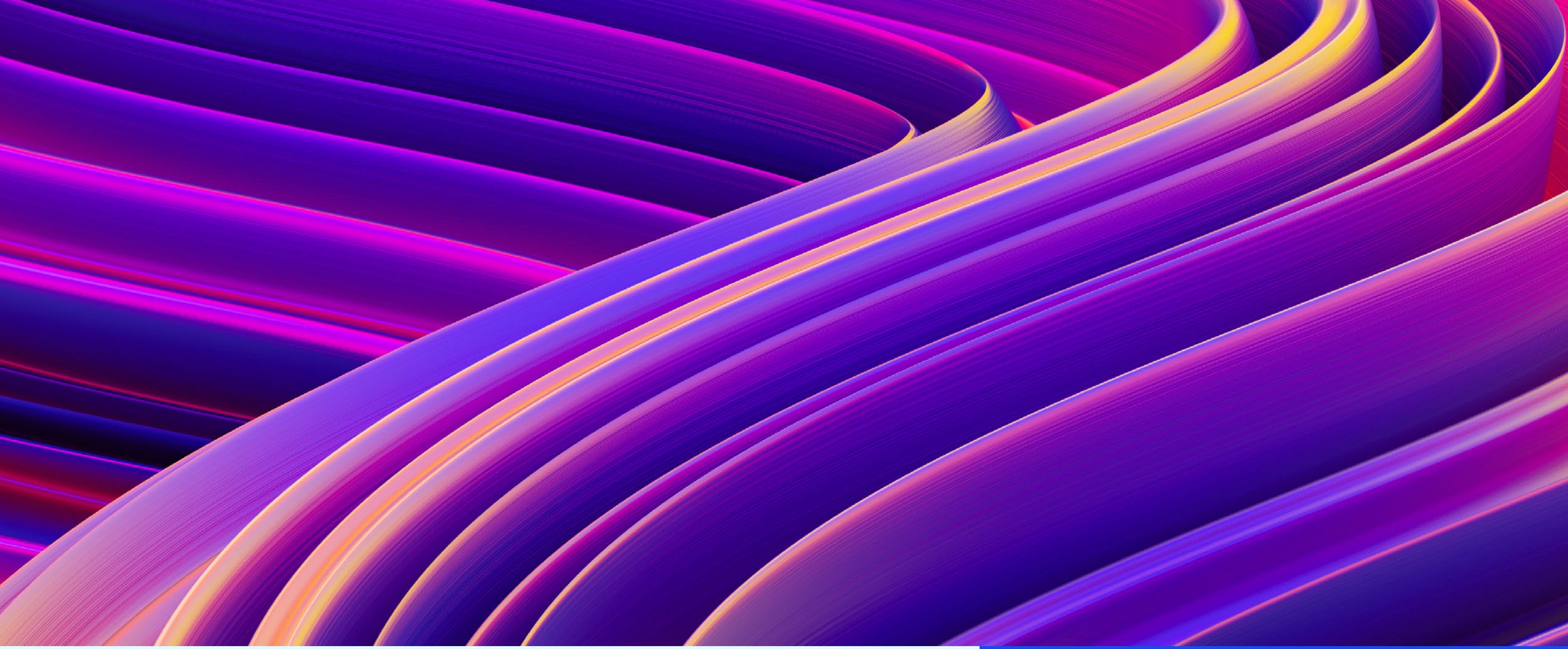
Small language models will likely play a role in the future of IA.

LLMs are getting the lion's share of the attention these days, but there is a case to be made the small language models (SLMs) can be impactful for IA teams. SLMs do not have the broad generalized capabilities that LLMs have. But they have many advantages. They can be trained and deployed at radically lower cost. They can be more easily fine-tuned to domain specific tasks. They are more sustainable in terms of energy consumption. And they are arguably more secure. While training an SLM is certainly not the first stop on your GenAI journey, the capability if worth bearing in mind for any innovative IA leader.

The future of AI is now.

The techniques described in this white paper are being used by the most innovative CAEs and their teams to perform their roles more effectively and efficiently. They are using GenAI to compare control matrices to industry frameworks, draft audit program guides, document process narratives, test internal controls, construct findings, and more. If your IA function is not focused on driving value with GenAI, the time to start is now.

This capability is being integrated into every major technology platform. The applications being generated using LLMs are rapidly generating business value. IA leaders have a professional imperative to embrace these technologies in order to retain credibility with their executives, meet the challenges of the profession, and to attract and retain talent.



KPMG can help.

KPMG has been recognized as the top provider for quality in AI advice and implementation services in Source's annual survey, Perceptions of Consulting in the US in 2024. Microsoft named KPMG the Supplier of the Year in 2024. Moreover, KPMG's Google Cloud Center of Excellence has been created to accelerate adoption of GenAI technologies. KPMG is the premier partner for GenAI in IA. We offer the following GenAI services for IA teams:

GenAI/Technology enablement strategy for IA or SOX teams: We can help you design clear goals and a roadmap to achieve you vision for a more technologically enabled IA function. We can also assist with building the business case for investment in this capability.

GenAI use case development: We can help you develop GenAI tools, prompt libraries and other enablement to get the most out of your GenAI strategy.

GenAI enabled audits: We conduct audits using GenAI capabilities and our global network of leading subject matter professionals.

AI governance audits: We can perform audits of your enterprise AI governance program.

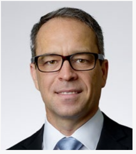
Wherever you are in your AI journey, our 15,000+ risk professionals can tailor our vast experiences, field-tested approaches, and advanced solutions to your organization, helping you navigate GenAI with confidence.

A note about GenAI risks.

All GenAI models and applications have inherent risks related to security, accuracy, safety, integrity, and other domains. These risks can be effectively managed with appropriate mitigations including contracts, configurations, training, and other internal controls. The KPMG Trusted AI framework provides a more comprehensive view of GenAI risks and risk management approaches. We encourage all IA teams to read our materials on Trusted AI to gain an understanding of these risks. For the purposes of this white paper, we assume that those interested in implementing the concepts herein will take precautions appropriate to their organization's GenAI governance policies.

Discover more insights at: visit.kpmg.us/FutureOfInternalAudit

Contact us



Charles King
US AI in Internal Audit Leader
Partner
KPMG LLP
E: crking@kpmg.com

Acknowledgments: Sreekar Krishna, Ph.D., Omid Yazdi, Max Hanson, Rachel Wagner-Kaiser, Ph.D., Swati Austin, and Diana Griffin all made invaluable contributions to this paper.