Collaborative AI-Based Multimodal Auditing: Integrating Foundation Models into Robotic Process Automation

Hanchi Gu, Marco Schreyer, Kevin C. Moffitt, Miklos A. Vasarhelyi

ABSTRACT

Audit procedures involve complex datasets across tabular, textual, and visual formats, but existing frameworks in auditing literature lack the flexibility to handle diverse procedures efficiently. This paper explores the feasibility of a collaborative AI-based multimodal auditing system that integrates foundation models into Robotic Process Automation (RPA) to automate audit processes. Experiments from different preset scenarios demonstrate that the latest publicly available foundation models have the potential to support such a system. In addition, the study demonstrates the importance of including non-routine audit procedures in RPA. The paper further introduces key terminologies related to generative AI to help accounting researchers better understand emerging technologies.

I INTRODUCTION

Auditing plays a critical role in ensuring financial transparency and regulatory compliance (IFAC 2009; AICPA 2021; PCAOB 2010b). Auditors are responsible for maintaining the accuracy and transparency of financial records and business operations. One of the primary duties centers on recording, reviewing, and analyzing various audit data. Auditors analyze a wide array of documents such as ledgers, journals, contracts, regulatory filings, and correspondence like emails and memos. The trend of digitization in accounting potentially enables auditors to reduce the time spent on collecting data (Krahel and Titera 2015), incorporating a more comprehensive range of data into the analysis for audits. While a more comprehensive inclusion of data in audit analysis can be advantageous, it also requires new forms of auditing to address the complexities introduced by big data. For example, the complexity of the big data in auditing can potentially cause information overload and impact audit judgment and decision making (Brown-Liburd, Issa, and Lombardi 2015). Existing literature has discussed how big data can bring potential transformations in the auditing domain (Brown-Liburd and Vasarhelyi 2015; Vasarhelyi, Kogan, and Tuttle 2015; Appelbaum, Kogan, and Vasarhelyi 2017). While the literature on Big Data discusses its Volume, Velocity, Veracity, and Variety, it overlooks a critical focus on Variety in depth. This aspect highlights the challenges posed by the diverse formats and types of data in audit analysis: the variety of data modalities.

In data science, “A modality is a category of data defined by how it is received, represented, and understood” (Vashney 2022). The variety of audit-related documents inevitably includes diverse data modalities. These include textual data, exemplified by regulatory filings, contracts, and agreements. Tabular representations include database tables and cost analysis spreadsheets. Visual materials comprise scanned documents, photographs, graphs, and charts. Computational scripts include various code scripts, database query languages, and interactions with external Application Programming Interfaces (APIs). Other modalities are also included. Auditors are responsible for analyzing the data in different ways. External auditors provide independent verification to enhance market confidence, while internal auditors offer in-depth evaluations of operational effectiveness (Johnstone, Gramling, and Rittenberg 2013, 1–3, 759–763). The absence of tools capable of integrating these diverse modalities can create difficulties for auditors in addressing the complexity of real-world audit procedures.

Advancements in data analytics technologies have now allowed for full population testing in audits. However, analyzing the entire population is often impractical due to time and resource constraints, especially with the increasing volume of operational data (Wang and Cuthbertson 2015). The complexity of multimodal data further contributes to the diversity and complexity of audit procedures. These procedures may involve investigations into anomalies, ad-hoc audits driven by regulatory inquiries, or the need to verify transactions across disparate systems. Unlike standardized, repetitive processes, these irregular procedures demand a high degree of flexibility and adaptability from auditors. These procedures often require different combinations of datasets from multiple modalities. As a result, they cannot be effectively addressed using software designed for specific processes. A key challenge with these procedures is that they frequently require processing and analyzing diverse datasets. To handle this complexity, next-generation automation tools to assist auditors must incorporate intelligent capabilities, enabling them to understand instructions for new procedures, apply reasoning to plan and solve problems, and organize diverse datasets to support audit procedures.

Robotic Process Automation (RPA) tools are software applications that automate repetitive, rule-based tasks by mimicking human interactions with digital systems. Existing RPA tools, such as UiPath and Blue Prism, excel at handling structured processes (Zhang, Thomas, and Vasarhelyi 2022) but struggle with the complexity and variability inherent in these irregular audit procedures. The reliance of current-generation RPA products on predefined rules and structured inputs limits their ability to adapt to the dynamic, multimodal nature of real-world audits. This gap highlights the need for more advanced solutions capable of integrating and processing diverse data modalities, enabling automation to extend beyond routine procedures and better assist auditors in addressing complex, evolving audit scenarios.

Adopting advanced artificial intelligence (AI) technologies offers a pathway to enhance existing RPA tools and bridge gaps in the current literature on audit automation. In the computer science domain, there is a phenomenon called the “modality gap”. This means that the model treats multimodal data such as images and texts as if they are inherently different and keeps them separate in its internal representation of knowledge (Liang, Zhang, Kwon, Yeung, and Zou 2022). With the recent evolution of AI, we witness a promising approach to bridging the “modality gap”, particularly with the emergence of foundation models like Large Language Models (LLMs).

Foundation models, which are pre-trained on extensive datasets as highlighted by Bommasani et al. (2021), offer adaptability across diverse audit procedures. The training mechanism of these models indicates that when trained on multimodal data, they are capable of effectively addressing multimodal problems, leading to high-performance results. Foundation models have the potential to mirror the human brain’s capacity to process multimodal data (Fei et al. 2022), thereby serving as a collaborative assist for auditors (Gu, Schreyer, Moffitt, and Vasarhelyi 2024).

The frontier of AI research has seen a recent trend toward general-purpose assistants capable of processing multiple data modalities (J. Li, D. Li, Savarese, and Hoi 2023). Nowadays, OpenAI's GPT[1], Google's Gemini[2], Meta's Llama[3], and Anthropic's Claude[4] all have multimodal abilities. Foundation models have the potential to follow natural language instructions and facilitate diverse multimodal audit procedures. While existing literature studies the application of deep learning in accounting (e.g. Sun and Vasarhelyi 2017; Sun 2019; Warren, Moffitt, and Byrnes 2015), research on the future applications of foundational models in auditing is still in its early stage (Gu et al. 2024; Föhr, Schreyer, Juppe, and Marten 2023; Eulerich and Wood 2023; Li and Vasarhelyi 2024). Our search, along with the review by Dong, Stratopoulos, and Wang (2024), indicates that no existing literature explores the multimodal capabilities of foundation models in auditing.

This study explores the possibility of extending existing RPA research to provide solutions for irregular and non-routine audit procedures, which have been overlooked in current auditing and accounting information system literature. These procedures often require auditors to handle diverse procedures with various data modalities, which existing RPA frameworks based on fixed workflows, struggle to manage effectively. Irregular and non-routine audit procedures demand that RPA tools go beyond automating predefined processes, requiring them to incorporate capabilities such as comprehension, reasoning, and adaptive problem-solving. The key challenge is enabling next-generation RPA tools to process diverse multimodal data and deliver intelligent solutions without excessive manual intervention.

This study contributes to literature by expanding existing RPA research to address a broader range of audit procedures. Second, it extends large language model research in auditing to include diverse data modalities. In the context of big data in accounting literature, it contributes

[1] https://openai.com/research
[2] https://gemini.google.com/
[3] https://llama.meta.com/
[4] https://claude.ai/chats

to enriching the discussion on data variety. Data variety is one of the four Vs of big data: Volume, Velocity, Variety, and Veracity. This aspect has not been well explored in accounting literature. This study further provides practical guidance on applying foundation models to real-world audit settings, demonstrating their ability to handle complex, non-routine procedures beyond the capabilities of traditional RPA tools. It also aims to inspire developers to design next-generation RPA tools.

The remainder of this paper includes six sections. Section II presents a comprehensive background review. It highlights the increasing complexity of audit data and procedures. It identifies key limitations in current RPA applications within auditing. It also outlines recent advancements in AI technologies. Together, these elements form the basis for exploring a collaborative AI-based multimodal auditing system. Section III proposes the concept of AI-based multimodal auditing systems, comparing them to traditional RPA and demonstrating their collaborative potential. Section IV introduces the overall experimental design, outlining the methodology used to evaluate popular models' performance across three levels of human-AI collaboration. Section V provides the detailed case study, including the specific scenarios and data used in the experiments. Section VI is the conclusion. Section VI analyzes the results, applying the experimental design from Section IV to the case scenarios described in Section V. Section VII is the conclusion.

II BACKGROUND

This section introduces the rising complexity of audit data and procedures, the development of Robotic Process Automation (RPA) in auditing, and its limitations. This section further introduces the need for advanced technologies that support auditors in handling diverse data modalities and non-routine procedures. In particular, it clarifies key terminology related to generative AI (GAI), which can be used interchangeably or inconsistently in accounting literature.

Complexity of Audit Data and Procedures

As businesses grow more complex and face evolving risks, audit data analytics (ADA) has shown new opportunities in audit procedures (Earley 2015). Big data significantly expands the scope of data, moving beyond financial to non-financial data, from structured to unstructured formats, and from internal to external sources. This shift challenges the audit profession's traditional comfort zone and technical capabilities (Alles and Gray 2015). To address the challenges of expanded audit data, researchers have explored automation, though their efforts primarily focus on repetitive and routine tasks (Huang and Vasarhelyi 2019). In real-world auditing, big auditors are taking on more non-routine tasks that go beyond repetitive tasks. The more varied the audit procedures, the more types of multimodal data auditors need to work with, ranging from spreadsheets and reports to code and images. This diversity in audit procedures naturally brings a broader mix of data, making it crucial to have flexible tools that can handle different formats and assist auditors in the audit process.

Diverse Audit Data

The diversity of audit data significantly complicates audit procedures because it involves managing various data types and formats from multiple sources. As a result, auditing requires advanced technologies to process and analyze these vast and complex datasets effectively. The digital transformation of companies has started an era where business processes and transactions are increasingly executed electronically, producing various data modalities. Consequently, the volume and variety of audit data have expanded significantly, becoming more complex and multifaceted. Auditors are responsible for evaluating the gathered audit evidence in audit procedures (AICPA 2011; PCAOB 2010a) for a wider range of data types. The data encompasses everything from databases to unstructured information found in emails, images, social media, and so on. A summary of common data modalities is summarized in Table 1. With different pre-training techniques, the modalities have further subcategories. For instance, code scripts incorporate Python, R, Java, C, and other formats. Audit firms will need to build skills in interpreting unstructured data and extensive datasets (Richins, Stapleton, Stratopoulos, and Wong 2017). Alles and Gray (2015) state that big data "remains a means towards an end and not, as the hype sometimes expresses it, as an end in itself." This means that the adoption of big data should enhance the value and efficiency of audit procedures for auditors, rather than adding unnecessary complexity or burdens to the process. When auditors are faced with audit data across multiple modalities, they can either manually process the data or rely on task-specific tools. Manual processing is undoubtedly labor-intensive, and task-specific tools come with significant costs associated with experts and programmers. If auditors depend on task-specific algorithms or tools to address audit procedures, the total number of combinations becomes significant (a detailed formula is shown in Appendix A). Task-specific systems require substantial investments in specialist expertise and training, necessitating domain experts and programmers to dedicate significant time to developing each specific

Data Modality	**Examples**
Textual Data	Regulatory filings, contracts, agreements
Tabular Data	Database tables, cost analysis spreadsheets
Visual Materials	Scanned documents, photographs, graphs, charts
Scripting and Code	Code scripts, database query languages, interactions with external APIs

Table 1. Examples of Audit-Related Data Modalities

function. However, if auditors had access to a simple solution capable of processing multimodal data within a single RPA tool, it would significantly reduce the cost of developing new algorithms and training auditors for new functions.

Non-routine Audit Procedures

Audit procedures vary in complexity, ranging from routine, repetitive processes to irregular and highly complex engagements that require deeper analysis and judgment. Irregular, unexpected, and non-routine audit procedures require RPA tools to move beyond simple rule-based automation. These procedures often demand procedure comprehension, contextual reasoning, and adaptive problem-solving. While each non-routine task may appear unique and isolated, collectively they can represent a significant portion of the audit workload.

Unlike repetitive processes, non-routine procedures call for flexibility, the ability to interpret ambiguous information, and the integration of diverse data sources. The challenge lies in equipping RPA to manage complex workflows, respond to new information in real time, and deliver intelligent solutions without requiring constant manual intervention or reprogramming. The audit standard does not specify the exact methods auditors must use to meet regulatory requirements. They only require auditors to form an expectation using reliable data from certain accounts and verify this to the recorded numbers (PCAOB 2016). Because of this lack of specificity in audit standards and the need for judgment in complex, irregular audits, the choice of analytical techniques is left to auditors and become a topic of growing debate with the rise of big data and automated financial reporting (Vasarhelyi, Kogan, and Tuttle 2015). Researchers have been integrating new technologies into audit procedures (Appelbaum, Kogan, and Vasarhelyi 2018). It is worth exploring how next-generation RPA tools can enhance automation and handle more diverse audit tasks.

Robotic Process Automation (RPA) in Auditing

Robotic Process Automation (RPA) is defined as an emerging technology that automates repetitive, rules-based procedures such as data entry, reconciliations, and confirmations, allowing auditors to focus on procedures requiring higher-level judgment (Huang and Vasarhelyi 2019). RPA works by mimicking human interactions with software systems, improving audit efficiency, accuracy, and cost-effectiveness (Zhang et al. 2022). Researchers has studied multiple aspects of its applications. RPA adoption in auditing faces challenges, including governance, integration with existing systems, and limitations in automating procedures requiring human oversight (Dahabiyeh and Mowafi 2023). As the technology evolves, its potential to transform the auditing profession is evident, but successful implementation requires careful management and oversight. More recent studies have shifted toward addressing governance and control frameworks to ensure sustainable and secure RPA implementations. Eulerich, Waddoups, Wagener, and Wood (2024) developed a comprehensive governance framework that includes four governance areas and 14 control requirements, validated through feedback from various RPA stakeholders. These advancements underscore the growing complexity of RPA deployments, highlighting the importance of governance, IT controls, and risk management (Zhang, Issa, Rozario, and Soegaard 2023). As the literature continues to develop, there is a clearer understanding of RPA's potential to enhance audit quality by automating repetitive procedures while also recognizing the need for frameworks that integrate human judgment, particularly in high-risk and complex areas.

The literature on RPA applications in accounting and auditing has evolved significantly over time. Early stages of research focused on automation techniques like Continuous Auditing (CA) and Continuous Monitoring (CM), which aimed to apply expert systems to real-time reporting (Vasarhelyi and Halper 1991). Later, RPA emerged as a distinct technology capable of performing structured, rule-based procedures, gaining widespread adoption in business process automation (Willcocks, Lacity, and Craig 2015). RPA is applied to audit procedures such as reconciliations, data entry, and internal control testing (Cooper, Holderness, Sorensen, and Wood 2019). However, most early frameworks focused on unattended RPA, where bots operate independently with minimal human intervention (Moffitt, Rozario, and Vasarhelyi 2018; Huang and Vasarhelyi 2019; Eulerich, Pawlowski, Waddoups, and Wood 2021). In response to the limitations of unattended RPA, the concept of Attended Process Automation (APA) emerged. APA involves active collaboration between auditors and bots, especially in procedures requiring real-time input or judgment, ensuring that automation complements rather than replaces human oversight (Zhang et al. 2022). This concept expanded the scope of RPA in auditing, as it allowed for the automation of more complex procedures that involve both human judgment and machine efficiency.

RPA is a broad concept that evolves as new technologies are integrated (Huang and Vasarhelyi 2019). While its core idea is automating procedures by mimicking human interactions with digital systems, RPA's ability expands with the adoption of advanced technologies. Understanding how new technologies have transformed RPA practices is crucial, as it provides prototypes for future case studies and archival research (O’Leary 2008). While each non-routine procedure may seem unique, collectively they often follow recurring patterns across audits, making them suitable candidates for RPA tools.

Regulators have also recognized the advancements in RPA. The IAASB (2023) highlights that RPA can automate tasks such as data preparation and transaction matching, but emphasizes the need for strong governance and auditor oversight (IAASB, 2023). Similarly, the PCAOB (Stein, 2023) stresses that technology-assisted analysis, including the use of RPA, must be carefully evaluated for relevance and reliability in audits (PCAOB, 2023). These developments underscore the clear need for further research in this field.

Advancements in Artificial Intelligence

AI advancements, especially in foundation models, enhance auditing tools and reshape auditors' roles. Auditors can collaborate with AI to easily automate tasks without requiring extensive technical or programming expertise, reducing reliance on programmers and enabling dynamic and flexible task automation.

Explanation of Terminologies Related to Generative AI

While existing literature contains various studies on AI following the recent popularity of ChatGPT (Dong et al. 2024), many accounting researchers still find certain terminologies confusing. Terminologies in accounting literature are often used interchangeably or inconsistently, resulting in conceptual confusion. This section offers explanations for some frequently used terminologies from computer science literature related to Generative AI.

Generative Artificial Intelligence (GAI) refers to a type of AI capable of creating new content, such as text, images, videos, music, and even code. Unlike traditional AI, which focuses on recognizing patterns or making predictions, GAI models generate entirely new data based on learned patterns (Bengesi et al. 2024). The current technological landscape remains diverse and continuously evolving, with popular approaches including Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), transformer-based models like GPT (Vaswani et al. 2017), and diffusion models (Ho et al. 2020). Popular models such as OpenAI's GPT-4, Google's Gemini, Meta's Llama, and Anthropic's Claude are specific models within the transformer-based model category.

Foundation models are a broad category encompassing large, pre-trained models capable of supporting a variety of downstream tasks (Bommasani et al. 2021). LLMs are a subset of foundation models. Other generative models, such as GANs and diffusion models, focus on creating new data but are not necessarily foundation models. Diffusion models, however, are increasingly scaling to foundation model status, while GANs typically remain more specialized and task-specific.

The term Large Language Model (LLM) originally refers to models designed to understand and generate natural language. The foundation models mainly focus on natural language in the beginning, which proposed the Transformer model and revolutionized natural language processing (Vaswani et al. 2017). However, LLMs are evolving beyond just language, incorporating other modalities like images, audio, and even video. Despite these broader capabilities, they are still referred to as LLMs, as language remains central to their functionality. Some key terminologies are summarized in Table 2.

Terminology	**Key Concept**
Generative AI	AI that generates new content from learned patterns
Foundation Models	Large, pre-trained, adaptable models for various tasks and modalities

Large Language Models	Language-focused models; now expanding to multimodal capabilities
Transformer-based Models	Models using the Transformer architecture; basis for most modern LLMs
GPT Series	OpenAI’s Transformer-based LLMs

Table 2. Key AI Terminologies and Concepts

To compare these terminologies, the term foundation model emphasizes the training mechanism by being pre-trained on large, diverse datasets, allowing them to be fine-tuned for a wide range of tasks (Bommasani et al. 2021). Unlike specialized models, foundation models are adaptable and can work across different domains (Li et al 2024), including multimodal data like language, vision, and more. Most modern LLMs are built using transformer-based architectures (Vaswani et al. 2017), but not all LLMs are transformer-based, as earlier models used Recurrent Neural Networks (RNNs) (Rumelhart, Hinton, and Williams 1986) and Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber 1997). In practical applications, GPT, Large Language Models (LLMs), and foundation models are often treated as synonymous, especially when LLMs are broadly defined to include tasks beyond just language processing. Most modern LLMs are built using the principles of foundation models. GPT series are the most famous LLMs. (the relationship among the terminologies is illustrated in Figure 1). Recent models have shown the abilities in arithmetic reasoning, commonsense reasoning, and natural language understanding (Yang et al. 2024).

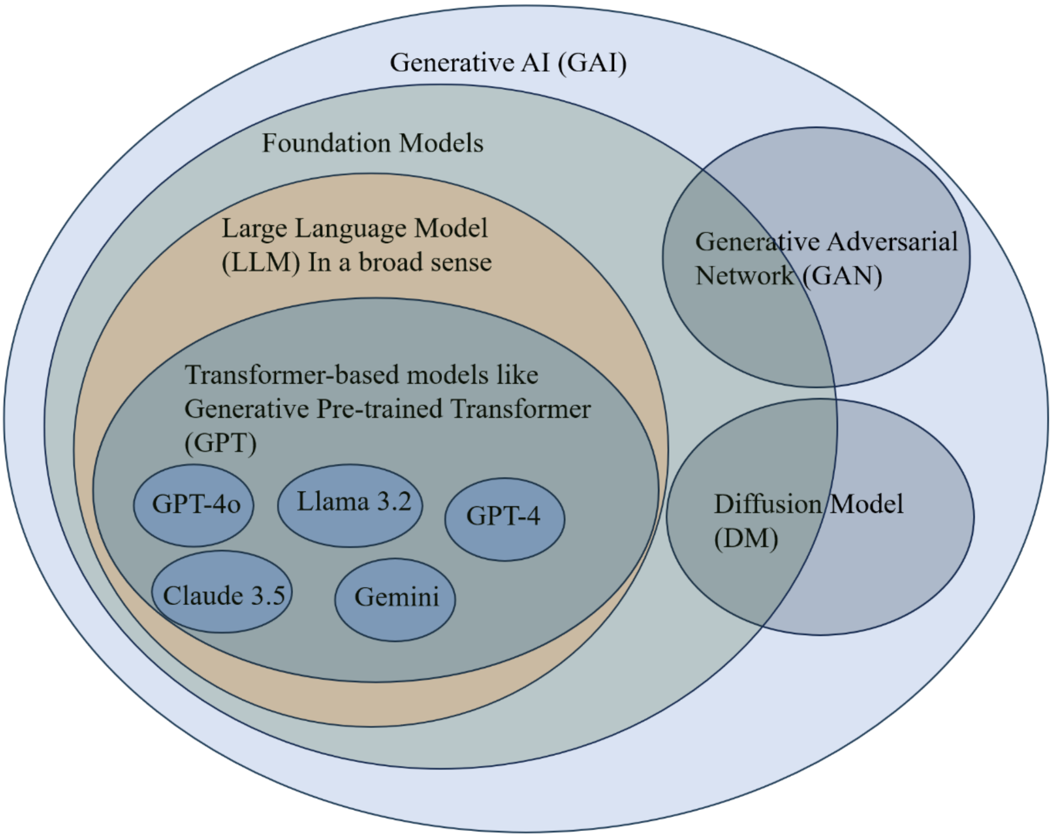

Generative AI (GAI)
Foundation Models
Large Language Model (LLM) In a broad sense
Transformer-based models like Generative Pre-trained Transformer (GPT)
GPT-4o
Llama 3.2
GPT-4
Claude 3.5
Gemini
Generative Adversarial Network (GAN)
Diffusion Model (DM)

Figure 1. Relationships of GAI-related Terminologies

Our demonstration of the key terminologies aims to help the accounting research community more clearly articulate and communicate concepts related to emerging AI technologies. This paper adopts the term foundation model to highlight that its ability to handle multimodal data comes from its pre-training mechanism on large and diverse datasets.

Multimodality in AI

The study of multimodality investigates how different data modalities, such as textual, visual, tabular, spatial, and audio data, work simultaneously to transmit information and generate understanding. Existing literature has developed many different multimodal models, such as the match of textual and visual data (Radford et al. 2021), text generation based on visual data (J. Li, D. Li, Xiong, and Hoi 2022; Li et al. 2023), complex images generation based on textual data (Ramesh et al. 2021). With a different focus, multimodal foundation models are also referred to as Multimodal Large Models (MLLMs), Visual-Language Models (VLMs), and other terms. In computer science, the trend involves expanding models to include additional modalities, such as code, audio, video, and 3D (Chen et al. 2021; Su et al. 2023; Zhao et al. 2023; Wang, Huang, Zhao, Zhang, and Zhao 2023; Hong et al. 2023).

Existing literature (Song, Li, and Li 2023) categorizes multimodal large models into four types. Multimodal Converter Models transform diverse multimodal inputs into formats that Large Language Models (LLMs) can process, enabling them to handle various data types. Multimodal Perceiver Models enhance LLMs' perceptual abilities by employing specialized modules to interpret information across different modalities cohesively. Tools-Learning Models extend LLMs' capabilities by enabling them to utilize external tools for data conversion and task execution, broadening their practical applications. Data-driven models focus on specifically training LLMs with targeted datasets to improve their performance and understanding in particular domains, such as medical imagery or point clouds, enriching their domain-specific applicability.

AI and Auditor Collaboration

The rapid advancement of AI has enabled models to understand diverse human instructions and interact in a user-friendly manner, moving closer to Artificial General Intelligence (AGI). AGI refers to systems that can learn, adapt, and perform across various tasks, similar to human intelligence, rather than excelling in only specific areas (Goertzel, 2014). As AI development edges closer to achieving AGI, studies on AI and human collaboration emerge in business research (Cao, Jiang, Wang, and Yang 2024; Anthony, Bechky, and Fayard 2023). The concept of AI co-piloted auditing has been proposed in auditing. AI co-piloted auditing refers to a collaborative approach where human auditors partner with advanced artificial intelligence to automate auditing tasks (Gu et al. 2024). In this partnership, the AI is capable of flexibly understanding audit data and comprehending auditors' instructions, completing diverse audit tasks. Current literature has not fully explored the potential of co-piloted auditing in handling data from multiple modalities. The potential for applying the co-piloted auditing concept to multimodal data remains a promising area of exploration.

Co-piloted auditing has the potential to improve existing RPA frameworks in auditing. In RPA literature, a key focus is on enabling humans to decide which tasks should be automated and which require human oversight. For example, in audit planning frameworks such as Attended Process Automation (APA), CPA firms determine which tasks will be automated and how the automation will be implemented prior to its execution (Zhang et al. 2022). A potential problem is that not all firms may fully understand how to effectively determine which tasks should be automated and which require human oversight. As a result, they may face challenges in making these decisions, leading to complex and time-consuming redesign processes during implementation. However, recent advancements in foundation models have taken this collaboration further. These models allow for a more seamless partnership between human auditors and machines. Human auditors no longer need to determine which tasks are best suited for automation precisely. Instead, they can work alongside AI systems that offer real-time assistance. This reduces the need for clear-cut divisions between human and machine tasks, as AI can dynamically support the auditors during their workflow, enhancing both efficiency and decision-making. These advancements in AI technology foster a more integrated and collaborative human-machine partnership in auditing.

III DEMONSTRATION OF THE SYSTEM

This section analyzes key considerations for integrating AI, particularly foundation models, into RPA. It presents the proposed three levels of AI-auditor collaboration. Finally, it summarizes how this approach differs from existing RPA frameworks.

Considerations for AI-Auditor Collaboration

As the study of audit data analysis evolves, there is an ongoing effort within the academic communities to develop new approaches and frameworks for effective audit data analytics (Krieger, Drews, and Velte 2021; Appelbaum, Kogan, and Vasarhelyi 2017; Kogan, Mayhew, and Vasarhelyi 2019). However, extant literature mainly works on single modalities (Duan, Vasarhelyi, Codesso, and Alzamil 2023; Gu, Dai, and Vasarhelyi 2023; Christ, Emett, Summers, and Wood 2021; Jans, Alles, and Vasarhelyi 2014; Li and Vasarhelyi 2024). Studies on combining multiple data modalities in auditing literature remain absent.

Multimodal foundation models can be a useful tool to assist auditors in interacting with foundation models to process audit data from various modalities in audit tasks. The model is expected to follow auditors' instructions regarding the description of the task, allowing the model to complete repetitive tasks with decent performance. Figure 2 provides a high-level illustration of the conceptual workflow of an AI-based multimodal auditing system. A multimodal audit system is characterized by several key elements that enhance auditing efficiency. First, models must be pre-trained on specific data types to handle future audit documents effectively. Second, auditors decide how the models collaborate with humans. Third, models can understand natural language, allowing auditors to interact naturally without extra training to complete audit procedures. Overall, such a system potentially improves the efficiency, accuracy, and user experience in integrating multimodal models into auditing processes. This section demonstrates the concept of AI-Based Multimodal Auditing, highlighting its highly collaborative and flexible nature, as well as its potential to reshape auditor workflows.

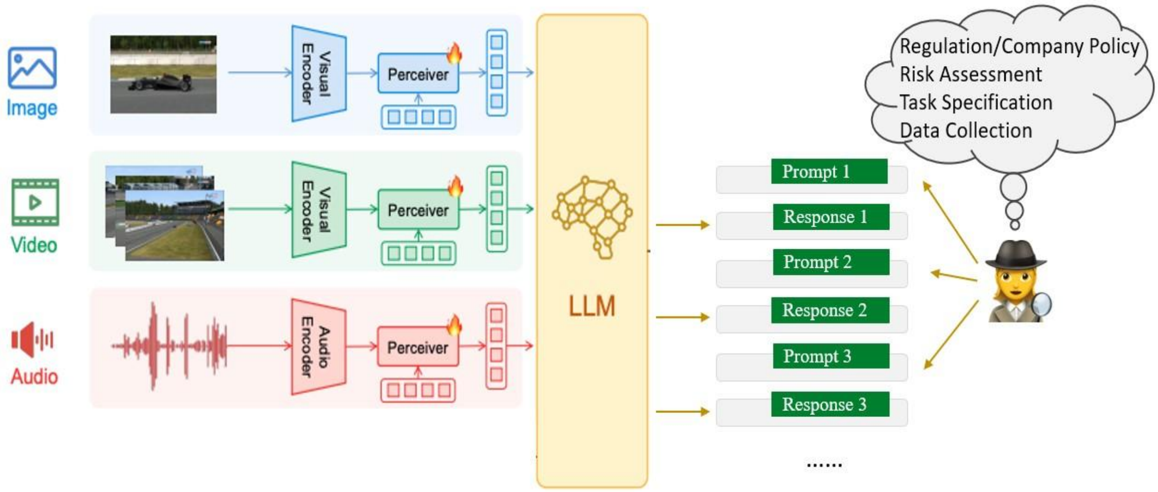


Figure 2. High-Level Overview of AI-based Multimodal Audit (adapted from Zhao et al. (2023) and Gu et al. (2024))

Although multimodal foundation models hold promise, they are not without limitations. LLMs may mimic training data without understanding, contributing to hallucinatory output to "cheat users with no perception" (Yao, Ning, Liu, Ning, and Yuan. 2023). The process of generalizing knowledge can inadvertently lead to memory distortions, introducing potential errors, as highlighted by (Yu, Zhang, Liang, Jiang, and Sabharwal. 2023). Moreover, the model also struggles to learn facts or concepts that are rarely mentioned in the training data (Kandpal, Deng, Roberts, Wallace, and Raffel 2023). Last but not least, multimodal tasks are often more challenging than tasks with single modalities because they require the integration and interpretation of diverse data types, which adds complexity to processing and analysis.

This complexity may cause the test result to be worse than the results in existing auditing literature on LLMs. Collectively, it is valuable to investigate whether foundation models can effectively handle multimodal audit tasks.

The collaboration between auditors and AI in audit procedures depends on two key factors: the capabilities of AI models and the complexity of audit tasks. The effectiveness of this collaboration is largely determined by the AI's ability to generalize knowledge accurately, minimize errors, and process unstructured or rare data. For simple, rule-based tasks, AI can operate with minimal oversight. For complex, multimodal tasks requiring deeper understanding, active auditor guidance is essential. As AI advances, it may handle more complex audits, but human oversight will remain crucial for critical judgment and nuanced analysis beyond AI's current capabilities.

Characteristics and Comparison

Collaborative AI-based multimodal auditing differs significantly from conventional RPA and Attended Process Automation (APA) in the auditing literature. The comparison highlights the distinct advantages of collaborative AI-based multimodal auditing, emphasizing its high degree of collaboration. Unlike conventional frameworks, AI-based multimodal auditing system allows human auditors to work closely with machines, functioning as intelligent assistants. The comparison is summarized in Table 3.

	Conventional RPA	**Attended Process Automation**	**Collaborative AI-Based Multimodal Auditing**
Summary	Automates rule-based, repetitive procedures using predefined workflows	Automates repetitive procedures with human collaboration for predefined functions	Leverages multimodal AI to create dynamic, collaborative automation as a partner
User Involvement	Minimal to none after deployment	Completes tasks assigned to human	Sends instructions in real time
Targeted Audit Procedures	Structured, rule-based processes	Structured tasks requiring some human interaction	Unstructured, non-routine scenarios
Workflow Adaptability	Requires human redesign for any changes or	Needs expert redesigns to adapt processes to new	Adaptive, capable of evolving and learning from new

	updates in workflows.	requirements or scenarios.	inputs without IT experts.
Implementation Approach	Fully automated, machine-driven execution	Parallel execution, with humans and machines working separately	Collaborative and adaptive, with frequent interactions

Table 3. A comparison of Conventional RPA, Attended Process Automation

Conventional RPA frameworks (Huang and Vasarhelyi 2019; Zhang et al. 2023) automates repetitive, rule-based tasks by following predefined workflows, making it ideal for handling structured processes with minimal variability. Once deployed, it requires little to no user involvement, allowing machines to take over execution. However, any changes or updates to workflows must be manually redesigned, limiting its adaptability to evolving audit needs.

Attended Process Automation (APA) (Zhang et al. 2022) builds on this by introducing human collaboration, enabling bots to assist users in real time. Automation handles repetitive, structured tasks. The workflow is designed and fixed for certain tasks. Users remain involved in overseeing and intervening in fixed steps. This concept extends earlier ideas introduced by Zhang (2019), who proposed Intelligent Process Automation (IPA) as a way to integrate RPA, AI, and other technologies to automate broader and more complex workflows. This approach allows for greater flexibility than conventional RPA but still relies on human experts to design and adjust workflows, which can slow down adaptability in rapidly changing environments.

Collaborative AI-based multimodal auditing represents the advancement of leveraging AI to manage complex, unstructured tasks through continuous collaboration with auditors. Unstructured tasks refer to irregular and infrequent tasks that lack predefined workflows, as experts have not designed automation processes for them. The key difference between the proposed system and existing RPA frameworks in auditing is its flexible and adaptive workflow, allowing it to handle unstructured, dynamic tasks through continuous collaboration with auditors. The system is designed to handle unstructured tasks, which are irregular and infrequent tasks lacking predefined workflows, making the workflow ad hoc and adaptable (shown in Figure 3). Unlike traditional systems that require redesigning for workflow updates, this system allows users to flexibly add new requirements, enabling adjustments without information technology experts. The proposed system fosters more active interaction between humans and AI, requiring both to understand and communicate. The collaborative and adaptive nature makes AI co-piloted auditing far more flexible and capable than conventional RPA and attended automation frameworks.

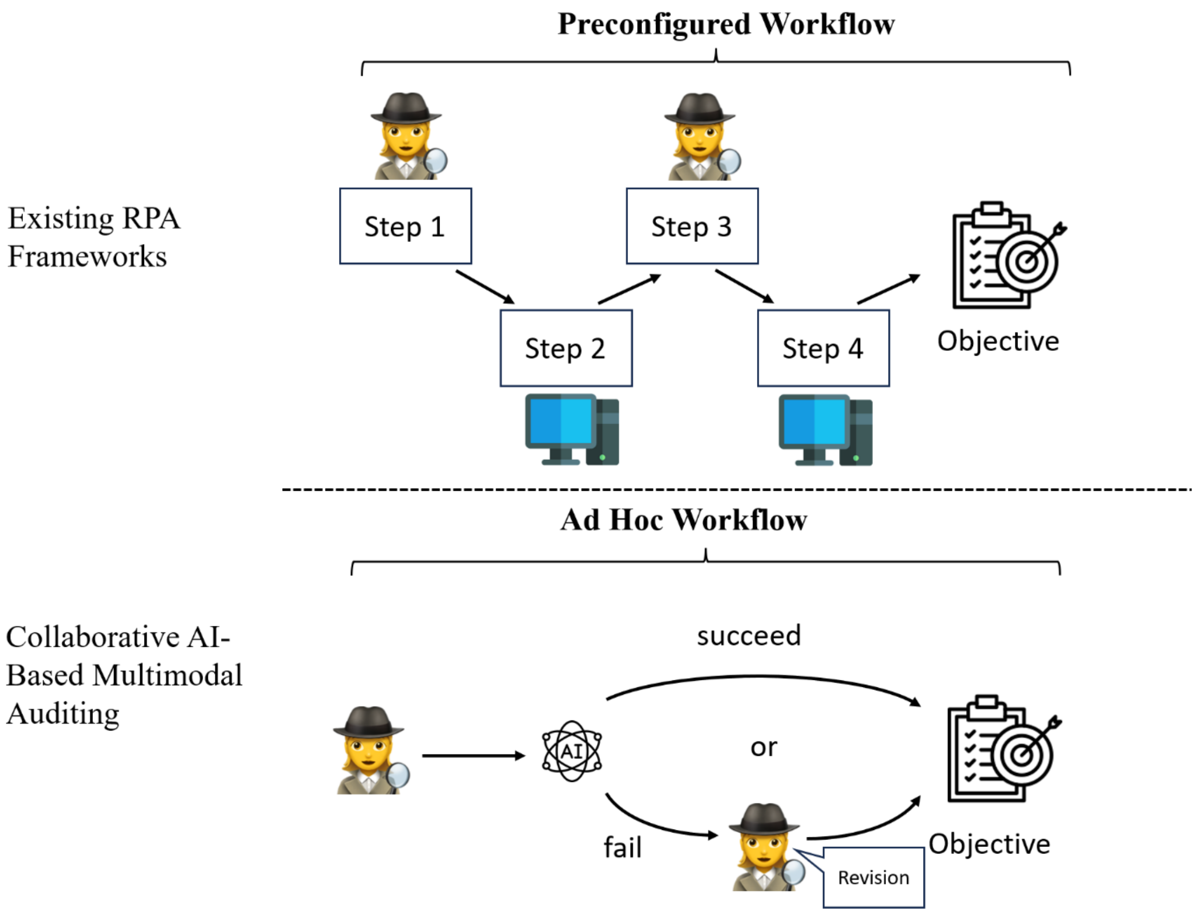


Figure 3. Collaborative AI-Based Multimodal Auditing Approach vs. Existing RPA Frameworks

Levels of AI-Auditor Collaboration

This study proposes three possible levels of collaboration for audit procedures (shown in Figure 4), reflecting varying degrees of AI involvement and auditor oversight. The first level focuses on automating specific functions, where auditors provide detailed instructions and well-crafted prompts to guide the AI in completing defined tasks. At this stage, the AI serves as a tool to enhance efficiency but still relies on human direction. The second level advances to automating entire audit procedures, with auditors setting objectives and supplying the necessary datasets. AI takes over the end-to-end process based on the audit procedure established by the auditor. The third and most autonomous level involves auditors feeding the available data into the system and allowing the AI to manage and explore the entire audit procedure independently without a clear objective. This progressive framework reflects the evolving role of AI in auditing, gradually shifting more responsibility to AI systems as their capabilities and reliability improve.

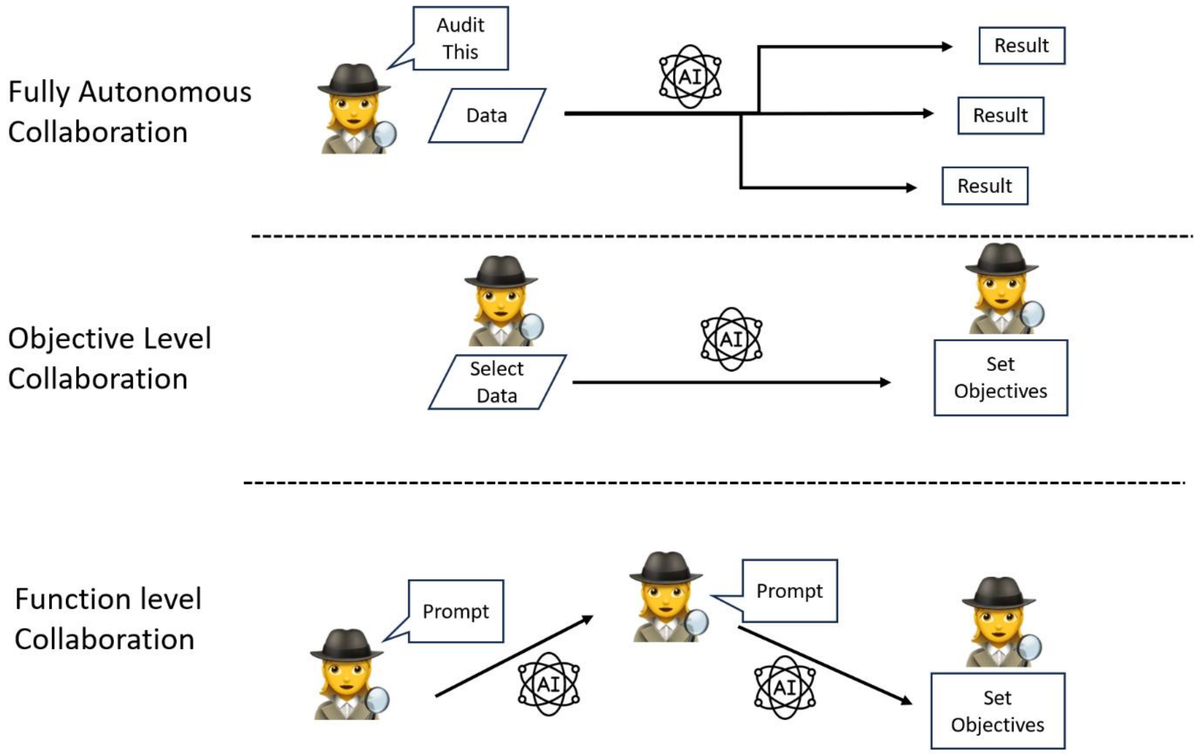


Figure 4. Three Levels of Collaboration

The three levels of collaboration illustrate how auditors can grant varying degrees of autonomy to AI models, enabling them to assist in audit procedures for specific functions, predefined objectives, or exploratory tasks. Generative AI models are continually evolving, as seen in the progression from OpenAI's GPT-2 to GPT-4o, with each version improving in accuracy and complexity. This development reflects the advancement in AI capabilities is an ongoing process.

This study serves as a normative exploration of AI-auditor collaboration, demonstrating and examining different levels of integration. While it provides a simplified workflow for future RPA practices, real-world implementations will likely be more complex due to factors such as regulatory constraints, data quality issues, human-AI interaction dynamics, and operational challenges. Existing literature also shows evidence that human-AI collaboration underperformed compared to the best of humans or AI alone, with performance losses in decision-making tasks but gains in creative tasks. Collaboration was beneficial when humans outperformed AI, but detrimental when AI outperformed humans (Vaccaro, Almaatouq and Malone 2024). Accounting researchers should conduct research in the future.

Despite their differences, all three levels function as collaborative AI-based multimodal auditing systems, integrating AI and auditors to enhance audit procedures with varying degrees of autonomy and human oversight. Nevertheless, this study can inspire researchers and RPA developers by providing insights into potential AI-driven RPA auditing systems and guiding future advancements in AI-assisted audit automation.

IV EXPERIMENT DESIGN

This section explains how different scenarios are designed to demonstrate the feasibility of a collaborative AI-based multimodal auditing system. This study will use experiments to demonstrate how an AI-based multimodal auditing system operates across three levels of collaboration: function-level, objective-level, and fully autonomous. This study adopts the experimental simulation approach in design science (Hevner, March, Park, and Ram. 2004), which showcases and evaluates the proposed three collaboration levels. A case study with synthetic audit data is created. The case study involves four data modalities: textual data, documentary images, figurative images, and tabular data. These experiments aim to assess the feasibility of such a system. We assume that foundation models will continue to evolve and demonstrate improved performance in the future. Therefore, if we can identify models that support the feasibility of a collaborative AI-based multimodal auditing system, it is reasonable to expect that more future models will also be capable of supporting such a system. Our goal is to collect evidence on whether top-ranked models[5] can support this type of collaboration. The tested models include Gork-3, GPT-4o, Gemini-2.0 Flash Thinking, and Deepseek-r1 based on their latest versions available as of March 2025.

The experiment follows a structured approach where we start with function level tasks, then move to more challenging ones, and finally test fully autonomous collaboration. If a model is unable to complete tasks at a more basic level, we discontinue testing it on more advanced tasks.

For function-level collaboration, the experiment assumes users can craft well-structured prompts, ensuring AI models perform tasks accurately and automate audit procedures. Prompt engineering is crucial for effectively programming large language models to perform complex

[5] This study refers to the overall ranking from the Chatbot Arena Overview (https://lmarena.ai/) as of March 2025. We select one model from each of the top four providers. In addition, we consider the accessibility of each model and include only those that offer unlimited access through their official websites.

functions and facilitate new interactions, enhancing their utility and interaction efficiency (White et al. 2023). The most general aspect of prompt engineering is the structure, often called the prompt template. Existing literature in computer science has discussed various templates to construct prompts, such as Chain of Thought (Wei et al. 2022), Tree of Thoughts (S. Yao et al. 2023), Graph of Thought (Y. Yao, Li, and Zhao 2023), Self-consistency Chain of Thought (Wang et al. 2022) and many other prompt engineering methods (Sahoo et al. 2024).

To align with our focus on irregular and non-routine tasks, the prompt is designed as a zero-shot prompt. While the field of prompt engineering is rich and complex, intricate prompt engineering skills can unnecessarily complicate their utilization. This paper serves as an exploration of applying multimodal foundation models to auditing practice with zero-shot learning, which does not require prior examples. Existing literature has shown applications of zero-shot learning prompts, such as financial documents (Hillebrand et al. 2023), recommendation systems (Hou et al. 2024), and legal documents (Savelka 2023). One challenge of embedding big data analytics in audit work is the development of relevant skills among auditors (Salijeni, Samsonova-Taddei, and Turley 2019). In real-world scenarios, auditors may lack existing examples of non-routine audit procedures. Our zero-shot prompt (shown in Appendix B.1), inspired by the copiloted auditing setting (Gu et al., 2024), strikes a balance between providing detailed instructions and ensuring ease of use, fostering effective interaction between auditors and AI models.

Objective-level collaboration and fully autonomous collaboration take advantage of foundation models' reasoning abilities. Existing literature demonstrates that AI models possess reasoning abilities, enabling them to analyze data, identify patterns, and draw conclusions with minimal human input (Moor et al. 2023; Qiao et al. 2022). These capabilities are fundamental to both collaboration levels, as they allow models to process audit data, follow objectives, and adapt to complex tasks, enhancing efficiency and accuracy in audit procedures.

For objective-level collaboration, we provide the model with audit data relevant to the task and set a clear objective, allowing the AI to handle the rest of the process (prompt template shown in Appendix B.2). In fully autonomous collaboration, we supply all available audit data and let the model independently complete the audit, leveraging its reasoning abilities to adapt without direct human involvement. This hands-on approach allows us to evaluate the model's performance across different levels of guidance.

V CASE STUDY

This section outlines the data and experimental design used in the study. A synthetic dataset is utilized to simulate the operational data environment of a synthetic company (documents are summarized in Appendix C). The data is generated independently from the experiments. This dataset encompasses purchase master data, policy documents, inventory daily records, and an array of visual documentation, including invoice, warehouse, and inventory images. A summary of the tasks is shown in Table 4. Documentary images are photographs or visuals that primarily aim to record and document. Figurative images represent real-world humans and objects. Although both documentary and figurative images are visual representations, they differ significantly in their purpose, content, and interpretation, making them different data modalities.

In a real-world setting, auditors need to consider diverse factors, such as audit documentation, risk, efficiency, and more detailed methods, to decide the implementation details. However, this paper alone cannot cover all the diverse details. This exploratory case study aims to demonstrate the feasibility of multimodal auditing through illustrative examples. For simplicity, this case study focuses on achieving accurate results and highlights important findings discovered during this process. This section consists of seven subsections. The first introduces six audit procedures involved, while the seventh summarizes the experiment results and insights.

	Documents Involved	Data Modality	Task description
1	Warehouse Inventory Transfer Supervision Policy, Warehouse image	Figurative Images, Text	Detect images that depict a manager wearing a white hard hat in the warehouse setting
2	Warehouse Safety Hard Hat Policy, Warehouse image	Figurative Image, Text	Identify images showing someone in the warehouse without a hard hat.
3	Inventory Accuracy Policy, Inventory image, Inventory Daily Records	Figurative Image, Text, Table	Match the number of stacks of disposable cups with the record.
4	Purchase Master Data, Purchasing Policy for One-Time Vendors	Table, Text	Identify suppliers or service providers with whom a business entity engages in a singular, non-recurring transaction
5	Purchase Master Data, Invoice Image, Invoice Filing Completeness Policy	Table, Documentary Images, Text	Identify purchase records in the Master Data lacking corresponding invoice images to ensure complete documentation
6	Purchase Master Data, Invoice Image, Alignment of Invoice Amounts Policy	Table, Documentary Images, Text	Identify purchase records in the Master Data that mismatch the amounts on the invoices

Table 4. Summary of Procedures in the Case Study

Manager Presence Check

In the context of warehouse operations, the transfer of a significant volume of inventory is a critical task that necessitates careful oversight. This procedure mandates the presence and supervision of a designated manager or supervisor to ensure that the transfer is conducted in an orderly, efficient, and secure manner. This oversight helps in verifying that the inventory is accurately accounted for, reducing the risk of errors, theft, or loss. From an auditor’s perspective, this practice is crucial as it aligns with internal control measures and compliance with related policies. Auditors evaluate such processes to ensure the integrity of financial records, safeguard assets, and assess the effectiveness of operational controls within an organization. By examining these practices, auditors can provide recommendations for enhancing controls, mitigating risks, and ensuring that the organization adheres to relevant laws, regulations, and industry standards. In this experiment, managers or supervisors are wearing white hard hats. The task is to identify the presence of a manager or designated supervisor, who is wearing a white hard hat in the images of a warehouse (Shown in Figure 5).

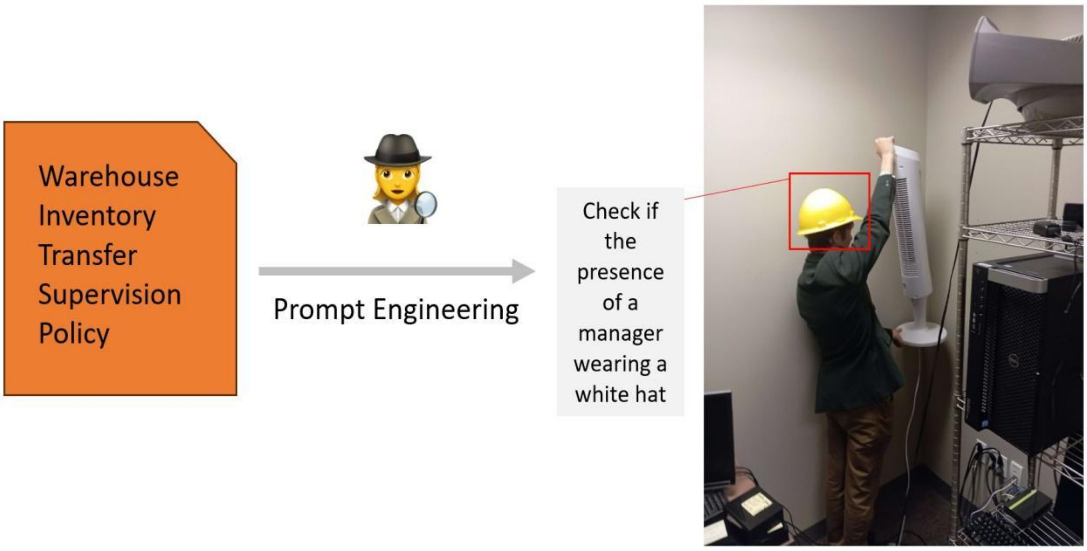

Warehouse Inventory Transfer Supervision Policy
Prompt Engineering
Check if the presence of a manager wearing a white hat

Figure 5. Manager Presence Check

Hard Hat Safety Verification

In the auditing process, the evaluation of safety protocols, such as the enforced use of hard hats in warehouses, serves a specific purpose. Auditors delve into the operational aspects of safety measures to ensure they align with the organization’s safety standards and regulatory requirements. Specifically, auditors may be responsive to the enforcement of hard hats, looking at how policies are communicated to and understood by employees, visitors, and contractors. This scrutiny is crucial for identifying risks, potential non-compliance issues, and areas for improvement in safety practices, ultimately ensuring that the organization’s safety measures are not only in place but are actively contributing to a safer work environment. In the Hard Hat

Compliance Check task, we apply the model to assess the use of hard hats among workers in warehouse images. Each image is examined to determine whether workers are adhering to safety regulations by wearing hard hats, with binary results generated to reflect compliance. Should any individual be identified without a hard hat, the system will display ‘Hard Hat Violation Detected’, signaling a breach of safety protocol. This method enables auditors to effectively enforce safety standards and identify lapses in compliance, ensuring a safer workplace environment (Shown in Figure 6).

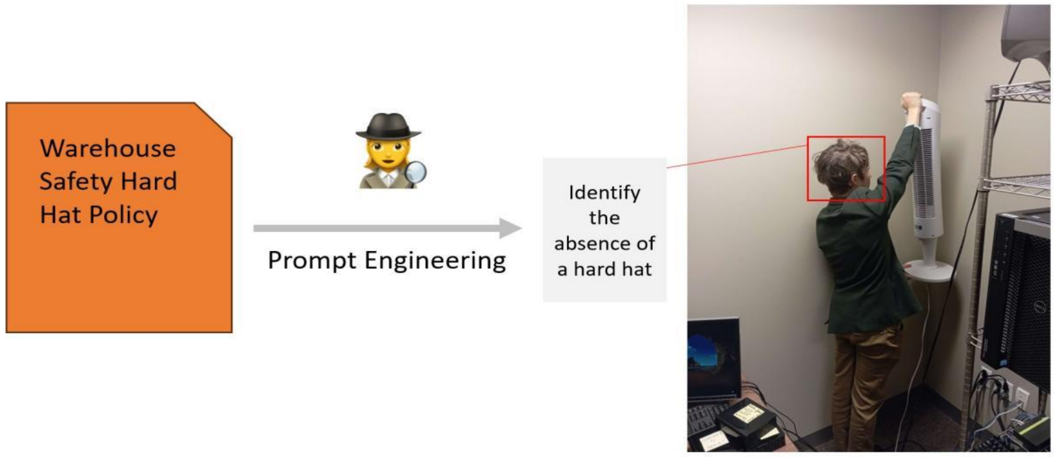
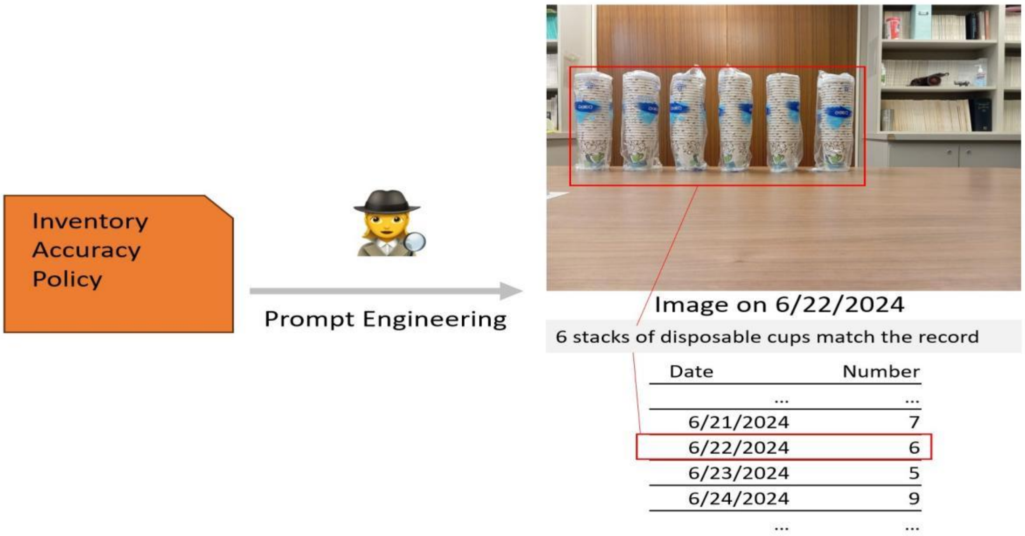

Warehouse Safety Hard Hat Policy
Prompt Engineering
Identify the absence of a hard hat

Figure 6. Hard Hat Safety Verification

Inventory Record Verification

Ensuring the accuracy of inventory records is vital for operational efficiency and financial integrity. Auditors examine the inventory management records to confirm they are accurate. Existing Literature has discussed similar problem in auditing (Christ et al. 2021). Such detailed assessment is crucial for identifying discrepancies, non-compliance issues, and potential areas for improvement in inventory management practices. The Inventory Accuracy Check task involves using a model to evaluate warehouse inventory records against the images of the inventory. In this experiment, the images of stacks of disposable cups are to match with the records in the inventory records. The model assists auditors in identifying any discrepancies

(Shown in Figure 7). In this experiment, we intentionally generate two inventory records with incorrect amounts and test whether the model can detect them.

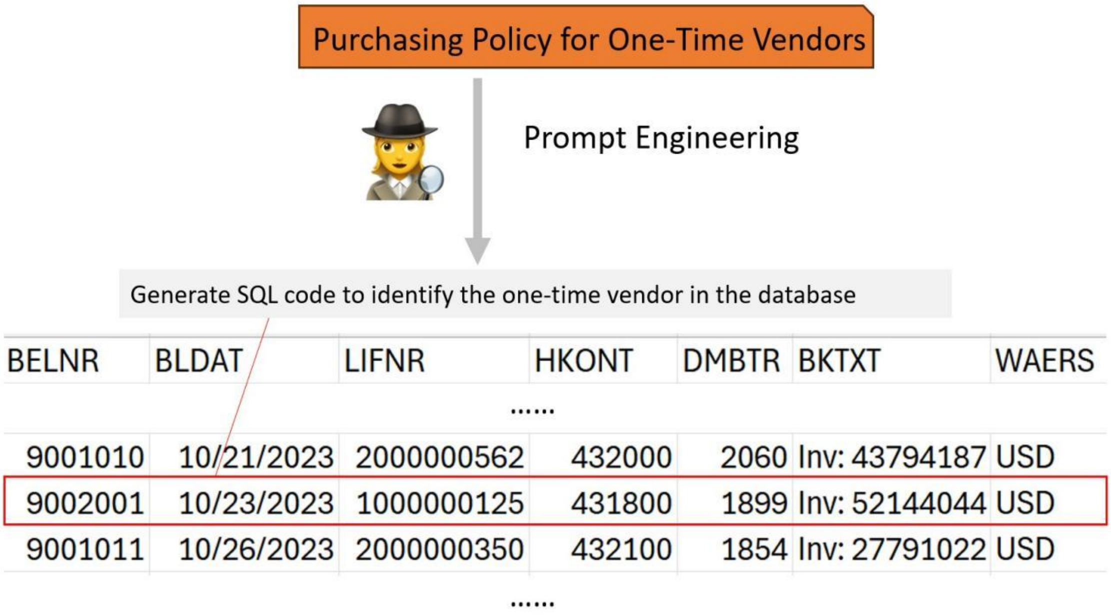

Inventory Accuracy Policy
Prompt Engineering
Image on 6/22/2024
6 stacks of disposable cups match the record

Date	Number
...	...
6/21/2024	7
6/22/2024	6
6/23/2024	5
6/24/2024	9
...	...

Figure 7. Inventory Record Verification

One-Time Vendors Risk

One-time vendors refer to suppliers or service providers with whom a business entity engages in a singular, non-recurring transaction, as opposed to establishing an ongoing relationship. Typically, these vendors are used for specific, unique purchases that do not necessitate continuous engagements. Unlike regular vendors, who may have contractual ties and established terms with a business, one-time vendors provide goods or services on a one-off basis, often without undergoing the rigorous vetting process that regular vendors usually face. In this experiment, we generate two one-time vendor records and insert them into the Purchase Master Data of the synthetic company. The goal is to produce an SQL code that can identify singular, non-recurring transactions, specifically transactions with unique values on the Account Number column (Shown in Figure 8).



Generate SQL code to identify the one-time vendor in the database

BELNR	BLDAT	LIFNR	HKONT	DMBTR	BKTXT	WAERS
			......			
9001010	10/21/2023	2000000562	432000	2060	Inv: 43794187	USD
9002001	10/23/2023	1000000125	431800	1899	Inv: 52144044	USD
9001011	10/26/2023	2000000350	432100	1854	Inv: 27791022	USD

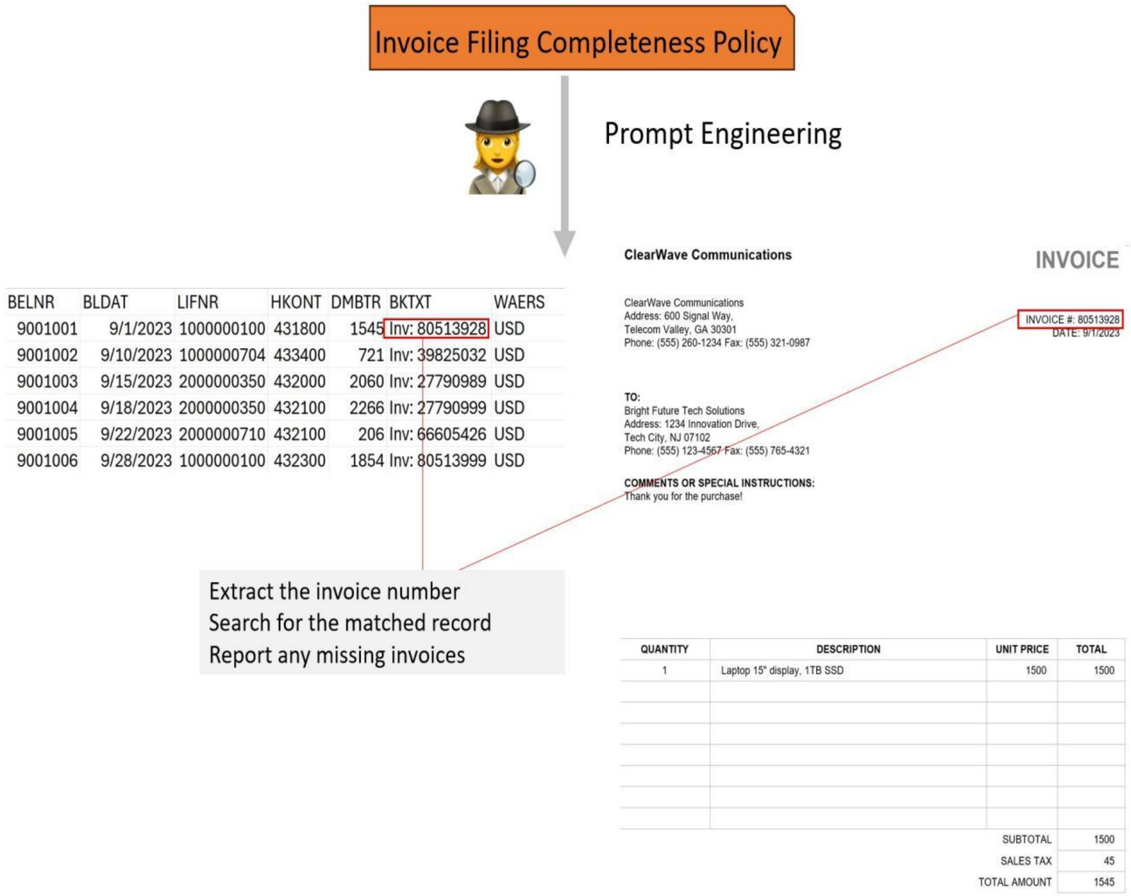
......

Figure 8. One-Time Vendors Risks

Invoice Filing Completeness

The documentation completeness check in the context of auditing involves the examination of existing records against a master list to identify any discrepancies or omissions. Specifically, when applied to invoice verification, this task entails comparing the collection of received invoices against the master data, which typically includes all expected invoices based on contractual agreements, purchase orders, or service deliveries. The primary goal of this audit task is to ensure that all financial transactions are accounted for, which helps in maintaining the integrity of financial statements. In addition, missing invoices can indicate issues such as unrecorded liabilities or potentially fraudulent activities. In this experiment, we delete two invoice images from the synthetic company's database. We start by extracting invoice numbers from scanned invoice images. Once extracted, these invoice numbers are stored in a dedicated table designed for easy integration with SQL queries. With the invoice numbers in place, our next step involves generating SQL code to cross-reference these numbers with the Purchase

Master Data, searching for purchase records that lack a corresponding invoice number (Shown in Figure 9).



Invoice Filing Completeness Policy

Prompt Engineering

BELNR	BLDAT	LIFNR	HKONT	DMBTR	BKTXT	WAERS
9001001	9/1/2023	1000000100	431800	1545	Inv: 80513928	USD
9001002	9/10/2023	1000000704	433400	721	Inv: 39825032	USD
9001003	9/15/2023	2000000350	432000	2060	Inv: 27790989	USD
9001004	9/18/2023	2000000350	432100	2266	Inv: 27790999	USD
9001005	9/22/2023	2000000710	432100	206	Inv: 66605426	USD
9001006	9/28/2023	1000000100	432300	1854	Inv: 80513999	USD

Extract the invoice number
Search for the matched record
Report any missing invoices

ClearWave Communications

INVOICE

ClearWave Communications
Address: 600 Signal Way,
Telecom Valley, GA 30301
Phone: (555) 260-1234 Fax: (555) 321-0987

INVOICE #: 80513928
DATE: 9/1/2023

TO:
Bright Future Tech Solutions
Address: 1234 Innovation Drive,
Tech City, NJ 07102
Phone: (555) 123-4567 Fax: (555) 765-4321

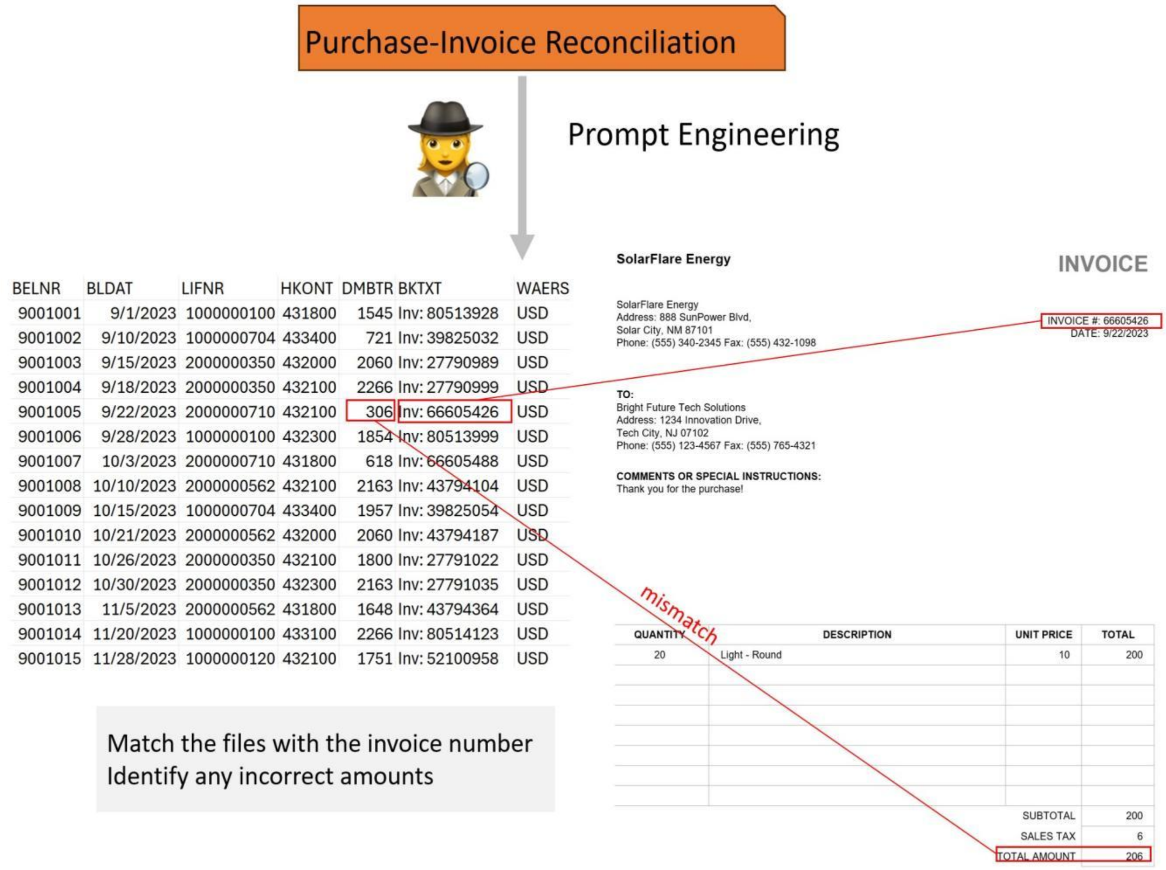
COMMENTS OR SPECIAL INSTRUCTIONS:
Thank you for the purchase!

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
1	Laptop 15" display, 1TB SSD	1500	1500
		SUBTOTAL	1500
		SALES TAX	45
		TOTAL AMOUNT	1545

Figure 9. Invoice Filing Completeness

Purchase-Invoice Reconciliation

In Purchase-Invoice Reconciliation, auditors are tasked with ensuring the accuracy of financial transactions by matching the monetary amounts listed on vendor invoices with those recorded in the organization’s purchase master data. This process involves a detailed examination of both the invoice details and the corresponding purchase master data. Discrepancies between these amounts can indicate errors or potential fraud, making this reconciliation process a critical component of financial auditing and internal controls. In this experiment, we make two records incorrect from the synthetic company’s purchase master data. The task is to assist in identifying purchase records in the Master Data with incorrect monetary amounts in the corresponding invoice images. The model exacts the invoice numbers and the total amount from the pictures. Then, SQL code is generated to compare the extracted information with existing records to identify any discrepancies (Shown in Figure 10).



BELNR	BLDAT	LIFNR	HKONT	DMBTR	BKTXT	WAERS
9001001	9/1/2023	1000000100	431800	1545	Inv: 80513928	USD
9001002	9/10/2023	1000000704	433400	721	Inv: 39825032	USD
9001003	9/15/2023	2000000350	432000	2060	Inv: 27790989	USD
9001004	9/18/2023	2000000350	432100	2266	Inv: 27790999	USD
9001005	9/22/2023	2000000710	432100	306	Inv: 66605426	USD
9001006	9/28/2023	1000000100	432300	1854	Inv: 80513999	USD
9001007	10/3/2023	2000000710	431800	618	Inv: 66605488	USD
9001008	10/10/2023	2000000562	432100	2163	Inv: 43794104	USD
9001009	10/15/2023	1000000704	433400	1957	Inv: 39825054	USD
9001010	10/21/2023	2000000562	432000	2060	Inv: 43794187	USD
9001011	10/26/2023	2000000350	432100	1800	Inv: 27791022	USD
9001012	10/30/2023	2000000350	432300	2163	Inv: 27791035	USD
9001013	11/5/2023	2000000562	431800	1648	Inv: 43794364	USD
9001014	11/20/2023	1000000100	433100	2266	Inv: 80514123	USD
9001015	11/28/2023	1000000120	432100	1751	Inv: 52100958	USD

Match the files with the invoice number
Identify any incorrect amounts

SolarFlare Energy

INVOICE

SolarFlare Energy
Address: 888 SunPower Blvd,
Solar City, NM 87101
Phone: (555) 340-2345 Fax: (555) 432-1098

INVOICE #: 66605426
DATE: 9/22/2023

TO:
Bright Future Tech Solutions
Address: 1234 Innovation Drive,
Tech City, NJ 07102
Phone: (555) 123-4567 Fax: (555) 765-4321

COMMENTS OR SPECIAL INSTRUCTIONS:
Thank you for the purchase!

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
20	Light - Round	10	200
		SUBTOTAL	200
		SALES TAX	6
		TOTAL AMOUNT	206

Figure 10. Purchase-Invoice Reconciliation

VI RESULT ANALYSIS

This section presents an analysis of the findings obtained after applying Gork-3, GPT-4o, Gemini 2.0 Flash Thinking, and Deepseek-R1 to the case study data. This section aims to examine the feasibility of integrating foundation models into RPA for multimodal audit procedures rather than determining which model performs better or providing detailed statistical comparisons.

In the function level experiments, four models were evaluated across six audit procedures using defined performance categories (summarized in Table 5).

- *Fully Performed* refers to cases where the model completed the audit procedure accurately and without errors.
- *Partially Performed* indicates the procedure was completed but with some errors.
- *No Support* signifies that the model was unable to handle certain data modalities required by the audit procedure.
- *Dysfunction* describes situations where the audit procedure could not be completed, and correct responses resembled random guesses.

DeepSeek-R1 lacks the capability to process figurative images, while Gemini 2.0 does not support tabular data. The results indicate that GPT-4o and Gork-3 demonstrated the highest performance, fully completing five out of six procedures, with both exhibiting dysfunction only in the Inventory Record Verification procedure. Gemini 2.0 partially completed two visually oriented procedures (18 out of 24 for Manager Presence Check, and 19 out of 24 for Hard Hat Safety Verification) but was unable to support the remaining ones due to its tabular data limitations. DeepSeek-R1 successfully performed only one procedure, with the rest either unsupported or dysfunctional. These results demonstrate the functional feasibility of a collaborative AI-based multimodal auditing system, as evidenced by the strong and consistent performance of two models, GPT-4o and Gork 3, across a majority of the audit procedures.

Audit Procedure	**Gork 3**	**GPT-4o**	**Gemini 2.0**	**DeepSeek r1**
Manager Presence Check	Fully Performed	Fully Performed	Partially Performed	No support
Hard Hat Safety Verification	Fully Performed	Fully Performed	Partially Performed	No support
Inventory Record Verification	Dysfunction	Dysfunction	No support	No support
One-Time Vendors Risk	Fully Performed	Fully Performed	No support	Fully Performed
Invoice Filing Completeness	Fully Performed	Fully Performed	No support	Dysfunction
Purchase-Invoice Reconciliation	Fully Performed	Fully Performed	No support	Dysfunction

Table 5. Experiment Results at Function Level

To assess the feasibility of objective-level collaboration, the two best-performing models, GPT-4o and Gork-3, were further evaluated across five audit procedures (summarized in Table 6). GPT-4o completed all tasks. Gork-3 fully completed four tasks but exhibited dysfunction in the Manager Presence Check, indicating a minor limitation in handling certain visual inputs. Overall, the results support the potential of using foundation models to enable collaborative multimodal auditing at the objective level.

Audit Procedure	**Gork 3**	**GPT-4o**
Manager Presence Check	Dysfunction	Fully Performed
Hard Hat Safety Verification	Fully Performed	Fully Performed
One-Time Vendors Risk	Fully Performed	Fully Performed
Invoice Filing Completeness	Fully Performed	Fully Performed
Purchase-Invoice Reconciliation	Fully Performed	Fully Performed

Table 6. Experiment Results at Objective Level

This study tested the feasibility of fully autonomous auditing using the GPT-4o model, which shows the best performance in previous tests. The system handled all tasks well and often went beyond simply checking rules. In the Manager Presence Check, it verified if a manager was present and wearing the required white hard hat. It also made useful observations about body posture, formal clothing, and engagement to suggest whether someone was truly supervising, even without clear documentation. In the Hard Hat Safety Verification, it confirmed hard hat usage and also paid attention to the surroundings. It questioned whether the person was in a warehouse or an office, which affects whether the policy applies. For the One-Time Vendor Risk task, the model checked G/L account use, flagged invoices over the $2,000 limit, and noted that most descriptions were missing or too vague. It also highlighted the absence of tax IDs, which the policy required. In the Invoice Filing Completeness task, it matched all purchase records with invoice images. It also flagged extra invoices that existed but were not recorded in the system, pointing out a potential oversight. In the Purchase-Invoice Reconciliation, the model compared invoice subtotals with SAP records and flagged mismatches, helping identify possible issues in recorded amounts. Overall, the system adapted well to different types of input and offered helpful insights. However, due to the limitations of our synthetic data, this study is unable to fully assess whether all of its long responses would hold up in real-world scenarios. Future studies using real operational data could expand on this work and more accurately evaluate the system's practical effectiveness.

In addition to the main results, this study also compares the latest GPT-4o model with an earlier version of GPT-4 (shown in Appendix D). The findings support the assumption that foundation models will continue to evolve and deliver improved performance over time.

VII CONCLUSION

This paper explores the integration of multimodal foundation models into Robotic Process Automation (RPA). Multimodal foundation models can collaborate with auditors in analyzing complex audit data and flexibly handling non-routine audit procedures. This paper demonstrates the potential of multimodal foundation models to enhance audit practices by effectively managing diverse data modalities. The paper discusses the idea of AI-based multimodal auditing systems with three different levels of AI-human collaboration. The paper compares the proposed approach with existing RPA application frameworks. A simulation using synthetic data demonstrates the feasibility of the designed methodology.

Collectively, the paper presents preliminary evidence that it is feasible to utilize multimodal foundation models to assist auditors in multimodal audit procedures. This study enhances the existing literature on applying ChatGPT to accounting by bridging the research gap related to multimodal data in accounting identified by Dong et al. (2024). This paper develops the concept of AI co-piloted auditing (Gu et al. 2024), which proposes that AI and auditors can collaborate like partners, to multimodal audit data. The study proposes different levels of collaboration in future AI-based auditing systems. The paper also enhances the existing literature on big data in auditing by exploring the variety of data modalities. This study benefits both audit professionals and academic researchers. The potential of multimodal foundation models can reduce manual labor and potentially increase the accuracy of audit outcomes, thereby boosting audit efficiency and effectiveness. This paper further contributes to the existing RPA literature by addressing the research gap in automating non-routine audit procedures.

Future studies can use real-world, large-scale data to better assess how the system performs under practical conditions. It will also be important to evaluate the quality of the model's responses by comparing them directly with expert judgments, especially when operating at a fully autonomous level. In addition, future research can explore developing foundation models into AI agents that combine foundation models with other technologies. For example, Retrieval-augmented generation (RAG) agents can enhance document retrieval capabilities (Setty, Jijo, Chung, and Vidra 2024; Gao et al. 2023), and Optical Character Recognition (OCR) can enhance image recognition abilities (Mori, Nishida, and Yamada 1999). It may mitigate the problem that some models cannot support certain data modalities. These directions may shape the next generation of robotic process automation in auditing.

USE OF GENERATIVE AI

During the preparation of this work, the author(s) used GPT-4o in order to polish the writing. GPT-4o, Gork-3, Gemini 2.0, and Deepseek r1 are used in the experiments. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCES

American Institute of Certified Public Accountants (AICPA). 2011. *AICPA Standard 315: Understanding the Entity and Its Environment and Assessing the Risks of Material Misstatement.*

American Institute of Certified Public Accountants (AICPA). 2021. *AU-C Section 200 Overall Objectives of the Independent Auditor and the Conduct of an Audit in Accordance With Generally Accepted Auditing Standards.* Available at https://us.aicpa.org/content/dam/aicpa/research/standards/auditattest/downloadabledocuments/au-c-00200.pdf.

Anthony, C., B. A. Bechky, and A. L. Fayard. 2023. "Collaborating" with AI: Taking a system view to explore the future of work. *Organization Science* 34(5): 1672-1694.

Appelbaum, D., A. Kogan, and M. A. Vasarhelyi. 2017. Big Data and analytics in the modern audit engagement: Research needs. *Auditing: A Journal of Practice & Theory* 36(4): 1–27.

Appelbaum, D., A. Kogan, and M. A. Vasarhelyi. 2018. Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics. *Journal of Accounting Literature* 40(1): 83-101.

Alles, M., and G. L. Gray. 2015. The pros and cons of using big data in auditing: a synthesis of the literature and a research agenda. *Awaiting Approval* 1: 37.

Bengesi, S., H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni. 2024. Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEe Access*.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258.*

Brown-Liburd, H., and M. A. Vasarhelyi. 2015. Big Data and Audit Evidence. *Journal of Emerging Technologies in Accounting* 12(1): 1–16.

Brown-Liburd, H., H. Issa, and D. Lombardi. 2015. Behavioral implications of Big Data's impact on audit judgment and decision making and future research directions. *Accounting horizons* 29(2): 451–468.

Cao, S., W. Jiang, J. Wang, and B. Yang. 2024. From man vs. machine to man+ machine: The art and AI of stock analyses. *Journal of Financial Economics* 160: 103910.

Chen, M., J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374.*

Christ, M. H., S. A. Emett, S. L. Summers, and D. A Wood. 2021. Prepare for takeoff: Improving asset measurement and audit quality with drone-enabled inventory audit procedures. *Review of accounting studies* 26(4): 1323–1343.

Cooper, L. A., D. K. Holderness, T. L. Sorensen, and D. A. Wood. 2019. Robotic process automation in public accounting. *Accounting Horizons* 33(4): 15-35.

Dahabiyeh, L., and O. Mowafi. 2023. Challenges of using RPA in auditing: A socio-technical systems approach. *Intelligent Systems in Accounting, Finance and Management* 30(1): 76–86.

Dong, M. M., T. C. Stratopoulos, and V. X. Wang. 2024. A Scoping Review of ChatGPT Research in Accounting and Finance. *International Journal of Accounting Information Systems* 55:100715.

Duan, H. K., M. A. Vasarhelyi, M. Codesso, and Z. Alzamil. 2023. Enhancing the government accounting information systems using social media information: An application of text mining and machine learning. *International Journal of Accounting Information Systems* 48:100600.

Earley, C. E. 2015. Data analytics in auditing: Opportunities and challenges. *Business horizons* 58(5): 493-500.

Eulerich, M., and D. A. Wood. 2023. A Demonstration of How ChatGPT Can be Used in the Internal Auditing Process. *Available at SSRN 4519583.*

Eulerich, M., J. Pawlowski, N. Waddoups, and D. A. Wood. 2021. Using robotic process automation in the internal audit function: Use cases and framework guidance for evaluation. *Contemporary Accounting Research*.

Eulerich, M., N. Waddoups, M. Wagener, and D. A. Wood. 2024. Development of a framework of key internal control and governance principles for robotic process automation. *Journal of Information Systems* 38(2): 29–49.

Fei, N., Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications* 13(1): 3094.

Föhr, T. L., M. Schreyer, T. A. Juppe, and K. U. Marten. 2023. Assuring sustainable futures: auditing sustainability reports using ai foundation models. *Available at SSRN 4502549.*

Gao, Y., Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997.*

Goertzel, B. 2014. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence* 5(1): 1.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27.

Gu, H., M. Schreyer, K. Moffitt, and M. A. Vasarhelyi. 2024. Artificial Intelligence Co-Piloted Auditing. *International Journal of Accounting Information Systems* 54:100698.

Gu, Y., J. Dai, and M. A. Vasarhelyi. 2023. Audit 4.0-based ESG assurance: An example of using satellite images on GHG emissions. *International Journal of Accounting Information Systems* 50: 100625.

Hevner, A. R., S. T. March, J. Park, and S. Ram. 2004. Design science in information systems research. *MIS quarterly,* 75-105.

Hillebrand, L., A. Berger, T. Deußer, T. Dilmaghani, M. Khaled, B. Kliem, R. Loitz, M. Pielka, D. Leonhard, C. Bauckhage, et al. 2023. Improving zero-shot text matching for financial auditing with large language models. In *Proceedings of the ACM Symposium on Document Engineering 2023:*1-4.

Ho, J., A. Jain, and P. Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33: 6840-6851.

Hochreiter, S., and J. Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8): 1735-1780.

Hong, Y., H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems 36*: 20482-20494.

Hou, Y., J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval,* 364-381.Springer.

Huang, F., and M. A. Vasarhelyi. 2019. Applying robotic process automation (RPA) in auditing: A framework. *International Journal of Accounting Information Systems* 35: 100433.

International Auditing and Assurance Standards Board (IAASB). 2023. Digital Technology Market Scan: Robotic Process Automation (RPA). Retrieved from https://www.iaasb.org/news-events/2023-01/iaasb-digital-technology-market-scan-robotic-process-automation

International Federation of Accountants (IFAC). 2009. *International Standards on Auditing (ISA) 200: Overall Objectives of the Independent Auditor and the Conduct of an Audit in Accordance with International Standards on Auditing.*

Jans, M., M. G. Alles, and M. A Vasarhelyi. 2014. A field study on the use of process mining of event logs as an analytical procedure in auditing. *The Accounting Review* 89(5): 1751–1773.

Johnstone, K. M., A. A. Gramling, and L. E. Rittenberg. 2013. *Auditing: A Risk-Based Approach to Conducting a Quality Audit.* 9th ed. Cengage Learning. https://nibmehub.com/opac-service/pdf/read/Auditing%20A%20Risk-Based%20Approach%20to%20Conducting%20a%20Quality%20Audit.pdf.

Kandpal, N., H. Deng, A. Roberts, E. Wallace, and C. Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning,* 15696–15707. PMLR.

Kogan, A., B. W. Mayhew, and M. A. Vasarhelyi. 2019. Audit data analytics research-An application of design science methodology. *Accounting Horizons* 33(3): 69-73.

Krahel, J. P., and W. R. Titera. 2015. Consequences of big data and formalization on accounting and auditing standards. *Accounting Horizons* 29(2): 409- 422.

Krieger, F., P. Drews, and P. Velte. 2021. Explaining the (non-) adoption of advanced data analytics in auditing: A process theory. *International Journal of Accounting Information Systems* 41:100511.

Li, C., Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao. 2024. Multimodal foundation models: From specialists to general purpose assistants. *Foundations and Trends® in Computer Graphics and Vision* 16(1-2): 1-214

Li, H., and M. A. Vasarhelyi. 2024. Applying large language models in accounting: A comparative analysis of different methodologies and off-the-shelf examples. *Journal of Emerging Technologies in Accounting* 21(2): 133-152.

Li, J., D. Li, C. Xiong, and S. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, J., D. Li, S. Savarese, and S. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730-19742. PMLR.

Liang, V. W., Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems* 35: 17612–17625.

Moffitt, K., A. Rozario, and M. A. Vasarhelyi. 2018. Robotic process automation for auditing. *Journal of Emerging Technologies in Accounting* 15(1): 1–10.

Moor, M., O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature 616*(7956): 259-265.

Mori, S., H. Nishida, and H. Yamada. 1999. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

O'Leary, D. E. 2008. Gartner's hype cycle and information system research issues. *International Journal of Accounting Information Systems 9*(4): 240-252.

Public Company Accounting Oversight Board (PCAOB). 2010a. *AS 1001: Responsibilities and Functions of the Independent Auditor.* Available at: https://pcaobus.org/oversight/standards/auditing-standards/analogo us-standards.

Public Company Accounting Oversight Board (PCAOB). 2010b. *AS 1105: Audit Evidence*. Available at https://pcaobus.org/oversight/standards/auditing-standards/details/AS1105.

Public Company Accounting Oversight Board (PCAOB). 2016. *AS 2305: Substantive audit procedures.* Available at https://pcaobus.org/oversight/standards/auditing-standards/details/AS2305.

Public Company Accounting Oversight Board (PCAOB). 2023. Algorithms, Audits, and the Auditor. Retrieved from https://pcaobus.org/news-events/speeches/speech-detail/algorithms-audits-and-the-auditor

Qiao, S., Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597.*

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning,* 8748–8763. PMLR.

Ramesh, A., M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning,* 8821–8831. PMLR.

Richins, G., A. Stapleton, T. C. Stratopoulos, and C. Wong. 2017. Big data analytics: opportunity or threat for the accounting profession? *Journal of information systems* 31 (3): 63–79.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. Learning representations by back-propagating errors. *nature* 323(6088): 533-536.

Sahoo, P., A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint arXiv:2402.07927.*

Salijeni, G., A. Samsonova-Taddei, and S. Turley. 2019. Big Data and changes in audit technology: contemplating a research agenda. *Accounting and business research* 49(1): 95-119.

Savelka, J. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law,* 447-451.

Setty, S., K. Jijo, E. Chung, and N. Vidra. 2024. Improving Retrieval for RAG based Question Answering Models on Financial Documents. *arXiv preprint arXiv:2404.07221.*

Song, S., X. Li, and S. Li, S. Zhao,　J. Yu,　J. Ma, X. Mao, W. Zhang. 2023. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint arXiv:2311.0759*4.

Su, Y., T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355.*

Sun, T. 2019. Applying Deep Learning to Audit Procedures: An Illustrative Framework. *Accounting Horizons* 33(3): 89–109.

Sun, T., and M. A. Vasarhelyi. 2017. Deep Learning and the Future of Auditing: How an Evolving Technology Could Transform Analysis and Improve Judgment. *CPA Journal* 87 (6).

Vaccaro, M., A. Almaatouq, and T. Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*: 1-11.

Vasarhelyi, M. A., and F. B. Halper. 1991. The continuous audit of online systems. *Auditing: A Journal of Practice & Theory* 10(1):110-125.

Vasarhelyi, M. A., A. Kogan, and B. M. Tuttle. 2015. Big data in accounting: An overview. *Accounting Horizons* 29(2): 381–396.

Vashney, K. R. 2022. *Trustworthy machine learning.* Independently published.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Wang, T., and R. Cuthbertson. 2015. Eight issues on audit data analytics we would like researched. *Journal of Information Systems* 29(1): 155-162.

Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171.*

Wang, Z., H. Huang, Y. Zhao, Z. Zhang, and Z. Zhao. 2023. Chat3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769.*

Warren J. D., K. C. Moffitt, and P. Byrnes. 2015. How Big Data will Change Accounting. *Accounting Horizons* 29(2): 397-407.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903.*

White, J., Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382.*

Willcocks, L., M. Lacity, and A. Craig. 2015. The IT function and robotic process automation. *MIS Quarterly Executive* 14(3): 173-188.

Yang, J., H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* 18(6): 1-32.

Yao, J. Y., K. P. Ning, Z. H. Liu, M. N. Ning, Y. Y. Liu, and L. Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469.*

Yao, S., D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems, 36: 11809-11822.*

Yao, Y., Z. Li, and H. Zhao. 2023. Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Large Language Models. *arXiv preprint arXiv:2305.16582.*

Yu, W., Z. Zhang, Z. Liang, M. Jiang, and A. Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002.*

Zhang, C. 2019. Intelligent process automation in audit. *Journal of emerging technologies in accounting* 16, 2: 69-88.

Zhang, C., H. Issa, A. Rozario, and J. S. Soegaard. 2023. Robotic process automation (RPA) implementation case studies in accounting: A beginning to end perspective. *Accounting Horizons* 37(1): 193–217.

Zhang, C., C. Thomas, and M. A. Vasarhelyi. 2022. Attended process automation in audit: A framework and demonstration. *Journal of Information Systems* 36(2), 101–124.

Zhao, Y., Z. Lin, D. Zhou, Z. Huang, J. Feng, and B. Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581.*

APPENDIX

A The Count of Tools Based on Task-specific Algorithms

The total number of task-specific algorithms needed is given by the formula:

$$A_{total} = \sum_{i=1}^{\sum_{k=2}^{n} \mathrm{C(n,k)}} M_i \times Ri \quad (1)$$

where $\mathrm{C(n,k)} = \frac{n!}{k!(n-k)!}$ is the total number of combinations of *n* items taken *k* at a time, excluding the e combinations with only one data type, and *Mi* represents the number of functions for the *i*-th combination. For each combination of modalities, different functions, such as classification and extraction, need to be designed separately. However, not all combinations of data modalities are equally important or necessary for audit procedures. Many combinations provide little incremental value, while others are essential for identifying risks and anomalies. To optimize this process, a binary selectivity factor R_i is introduced for each combination. R_i is a binary value of 0 or 1, selectively includes or excludes each combination and function based on its relevance.

B Prompt Template

B.1 Template 1

"Your task is to assist in [general description of the audit procedure]. Utilize actions such as [more specific description of the tasks, such as generating SQL code, Python code, Extracting, etc.] on the provided [description of the input data] with attributes like [sample input data]. The goal is to produce a [description of the expected output data]. You should [step-by-step instructions on how to complete the audit-related task. This part can be separated into more than one prompt for step-by-step execution]."

B.2 Template 2

"[Describe the objective of the audit procedure]"
Upload the data and related internal control policy

C Synthetic Company Details

Below is a summary table of the various documents within Bright Future Tech Solutions:

Document Type	Description	Quantity
Invoice Images	Images of Invoices	20
Warehouse Images	Images of people in warehouse	24
Inventory Images	Images of stacks of disposable cups	15
Purchase Master Data	20 Records of purchase details	1
Inventory Daily Records	15 Records of Daily Inventory Amount	1
Policies for Audit Tasks	Guidelines and rules for conducting audit tasks	6

Table C.1: Summary of Documents in Bright Future Tech Solutions

In the case study, this study intentionally changes the sample to create some scenarios, including incorrect inventory records, one-time vendors, missing invoices, and incorrect amounts.

D Test Results on old version of GPT-4

This section reports the test results in March 2024 based on GPT-4 models. As a comparison to our latest test in March 2025. Our findings indicate five of the six tasks can be solved successfully with several iterations, except the Inventory Record Verification task. First, the information load can affect the model performance. For tasks involving image data, one photo is added at a time, and this process is repeated, as the GPT-4 model struggles to handle multiple photos effectively for these two tasks. Second, variability in actions can be an important factor. For Manager Presence Check, the model sometimes failed because the current version of the GPT-4 model can address multimodal problems either through its inherent multimodal capabilities or by leveraging Python packages. If the model utilizes a Python package with poor performance, the model can fail. Additionally, the first trial of the Purchase-Invoice Reconciliation did not succeed in all tasks. The same prompt was used to repeat the tasks for the failed extractions. The second trial successfully completed the task, demonstrating the inherent randomness of LLMs.

However, the model cannot complete Inventory Record Verification because the current version of the GPT-4 model fails to count the number of stacks of disposable cups. Even if we modify the implementation details or adopt Template 2 to enhance the model's comprehension of the instructions, the model still fails to complete the task. Since similar problems have been successfully addressed in existing literature (Christ et al. 2021), current algorithms should be capable of resolving this issue. We anticipate that future foundation models could refine the counting function and potentially overcome these challenges. In conclusion, the results from the case study demonstrate the potential and feasibility of applying multimodal foundation models, like GPT-4, to auditing practices. The study shows that auditors can adapt implementation details to successfully complete most tasks. However, current models still have limitations, particularly in functions like counting items in specific scenarios, indicating areas for future improvement.