Exploration on Application of Large Language Models in Unstructured Audit Materials

Weiyi Li

Business School, University of Auckland, Auckland, 1010, New Zealand

Abstract. Research on the application of large language models (LLMs) and natural language processing (NLP) in unstructured audit data. Traditionally, auditors have used only the quantitative financial data, and unstructured data such as emails, contracts and meeting minutes have been difficult to process on a large scale, thus increasing audit risk. Newer LLMs can be used to combine text and numbers in order to strengthen the risk assessment and financial statement analysis work. Review of the theory and practice of AI application in unstructured audit evidence. Research has been conducted on how to use large language models (LLMs) for automated compliance checks, sentiment analysis and anomaly detection of narrative context combined with quantitative disclosures. In addition, there are technical and governance problems with the above models, such as algorithmic bias, data privacy, and the requirement for continuous monitoring. Although LLMs have enhanced the extraction and analysis of unstructured data, a strong governance system and an immutable audit trail are needed for their application to ensure professional scepticism and regulatory compliance in the new era of digital audits, as mentioned in this paper.

Keywords: Large Language Models, Unstructured Data, Audit Evidence, Natural Language Processing, Risk Assessment, Continuous Auditing.

1. Introduction

The audit profession has been organised based on the structure of quantitative financial data for a long time. The foundation of the traditional audit evidence includes the ledger, trial balance and standardised financial statements. However, today's enterprises have produced a large amount of unstructured data, such as emails, board meeting minutes, complex legal contracts and social media posts. Although this narrative information contains necessary background and reasons for management's decisions, auditors have not had access to all such data systematically throughout the company in the past.

Manual review of the unstructured data is computationally expensive and limited by human sampling, thus increasing audit risk. Therefore, the supervision of financial reporting often does not use the textual information that explains the changes in figures. With the rapid development of artificial intelligence (AI) in recent years, natural language processing (NLP) and large language models (LLMs) have become increasingly sophisticated. Deep learning models have been introduced to understand context, sentiment and semantic relations in previous text data, and now machine-readable indicators have been extracted.

Narrative and numerical data are combined to perform continuous, all-weather monitoring by the system now. LLMs can automate compliance mapping and behaviour anomaly detection, and extract key contract provisions more accurately than previously possible. However, this change in technology is also imperfect. Generative AI is now being applied to high-risk audit scenarios, and some new risks have emerged, such as model hallucinations, bias in training data, and complicated data governance issues.

This paper investigates a strategy for applying LLMs to unstructured audit data. Section 2 introduces the theory of analysis for accounting-related unstructured text data. Section 3 describes how LLMs improve audit risk assessment and evidence collection functionally. Section 4: Ethical, Technical and Governance Issues in Algorithmic Auditing. Section 5 is the operating model for ongoing observation and compliance. Section 6 is the conclusion, and the following will be discussed.

2. The Theoretical Evolution of Text Analytics in Auditing

The foundation for applying textual data in the audit process is that the numerical disclosures resulted from human judgement, negotiation and other factors. Fisher et al. [2] have compiled the initial studies on natural language processing in accounting and shown that text analysis can be applied to evaluate the quality of internal controls and management intentions. Before the spread of deep learning models, most text analysis was done by means of rule-based keyword dictionaries and simple frequency statistics, and it could not handle irony, negation or other complex financial situations.

Transformers have been developed to create more advanced natural language processing (NLP) systems and large language models (LLMs) that can produce continuous vector representations of unstructured data. Dong and others [10] have recently focused on classification, summarisation, text generation and analytical support for financial decision-making in their research on ChatGPT and LLMs in accounting and finance. This ability addresses a deficiency in the previous risk model; that is, using only numerical factors cannot detect financial distress or subtle misstatements in an economy downturn.

Unstructured text can be viewed as a type of formalised audit evidence, and thus the whole-picture risk assessment system will be more comprehensive. Amoatey [1] believed that although text has clear analytical benefits, the absence of an organised system for applying NLP throughout the audit cycle has prevented its use to date. A design science approach is needed to integrate these technologies into the essential links of planning, assessment of internal control, and substantive testing to ensure that text analytics enhance auditors' judgement rather than replacing it. If the unstructured data can be mapped reasonably to the financial assertions, then it will no longer be an administrative burden but rather necessary support for compliance assurance.

In addition, this theoretical change also expands the scope of external audits historically. Auditors can now process all kinds of unstructured data uniformly, so the demand for fraud detection and going-concern assessments will also rise. Therefore, the theoretical requirement has changed: if technology can fully analyze management communication and external market information, excluding such analysis is no longer justifiable under the new audit standards.

In addition to extracting latent features, the theoretical basis of text analysis also relates to the problem of information asymmetry in corporate governance. Management is more familiar with the actual circumstances of operation and therefore more so than external auditors. In the past, such an asymmetry was addressed by testing the internal controls over financial reporting. However, with an increase in the scale of organisations, the deficiencies of sampling structured data become more evident. Sophisticated language models are used to analyse unstructured data, such as the qualitative disclosures in management discussion and analysis (MD&A) reports or the syntactic structure of internal communications, and an independent secondary mechanism for assessing the veracity of management's statements is thus established. The two-tier verification system will automatically reduce information asymmetry and help the auditor process the large amounts of data produced by modern enterprises. Therefore, the theoretical basis for employing unstructured data analysis in an audit has shifted from being an option to one that must now be realised to achieve reasonable assurance under complex audit circumstances.

3. Functional Mechanisms of LLMs in Audit Evidence and Risk Assessment

LLMs' main functional benefits for auditing are to contextualise and extract specific financial data from large volumes of unstructured text. Li et al. [7] have built a new system that can automatically extract the main financial indicators from unstructured PDF reports accurately using LLMs. Prompt engineering and text mining can be employed to map the information in unstructured, varied-format corporate reports to a database structure and avoid tedious manual verification and tracing work for thousands of hours.

LLMs can be used to compare the content of a risk assessment with numerical disclosures in the area of risk assessment. Kim et al. [3] have investigated the application of artificial intelligence in enhancing the supervision of financial reporting, and by integrating narrative context into predictive models, they have significantly improved the detection rate for directional changes in earnings and high-risk misstatements. The LLM can find any changes in the mood of the management team or evasive language that is inconsistent with the optimism shown in the numerical forecast during the conference call. Both modes will be used to raise the audit risk at the same time.

A deep-learning model can be used to construct an audit plan automatically according to unstructured compliance requirements. Föhr et al. [6] have shown that deep learning models are good at information extraction, and unstructured data can be turned into machine-readable form for risk-based auditing. For example, an LLM can read a difficult lease contract to find specific restrictions and obligations, check them against the recognized lease liabilities in the balance sheet at the same time, etc. Thus, the auditor will no longer need to collect data but only analyse it.

European securities and markets authority [5] has also determined that LLMs can support all links in research and decision-making through analysis of large volumes of regulatory disclosures, legal documents, and so on. By compressing the long chain of operational communications into a decision-support context, these models can continuously inform engagement teams of changes in a client's risk situation in real time and are thus much more efficient than periodic sampling.

4. Governance, Bias, and the Ethical Dimensions of Algorithmic Auditing

Large language models (LLMs) are good at analysis, but they also have certain problems for regulating the high-risk application of auditing. The first type of risk is that deep neural networks are opaque and prone to hallucination; thus, they may generate false financial information. Fotoh and Mugwira [8] have studied the use of LLMs in external audits and proposed that due to ethical problems in applying generative models, stringent verification measures must be taken. An auditor cannot use the output of an algorithm to support a modified audit opinion without revealing the reasons for that output.

Algorithmic bias is another type of risk. LLMs have inherited the biases of the training data. If a model is used to identify abnormalities in employee expense reports or internal emails, some linguistic styles or demographic groups may be incorrectly identified as suspicious. Protiviti [4] thinks that, in using natural language processing for internal audits, caution must be exercised to prevent any kind of discrimination caused by risk-scoring algorithms under the pretense of objectivity and automation. Reduce the above risks by fine-tuning a domain-specific model based on the standard accounting taxonomy and do not use general web data.

Protect the privacy and security of data. Unstructured data may contain personal information and other private or confidential data of a company. Kokina and others [11] have listed the following problems in the application of AI to auditing: lack of transparency, explainability, data privacy, bias, reliability and the risk of excessive reliance on auditors. When auditors have the LLM process client emails and legal documents for analysis, they need to use a secure, isolated architecture (enterprise Retrieval-Augmented Generation) to ensure that the client data does not contaminate the training data of public models.

Given the above risks, the former standards for software procurement are no longer suitable. The Institute of Internal Auditors [9] has built an all-encompassing AI Audit System to oversee the implementation. The platform will help organisations build an all-encompassing index of their AI systems, create data-access segmentation boundaries, and implement strict cybersecurity measures for all repositories of processed unstructured data. Without the above governance, the efficiency gains of LLMs would be cancelled out by regulatory and reputational risks.

In addition to algorithmic bias and hallucination, there is also a problem of explainability in audit documentation due to the use of LLMs. The black-box characteristics of deep learning models are inconsistent with the demands of audit standards for thoroughly documented, reproducible and comprehensible audit procedures carried out by experienced auditors who have no prior contact with the audit. If the neural network identifies a complicated leasing arrangement as a high-risk anomaly, the engagement team cannot simply record 'algorithmic determination' as the reason for not performing further substantive tests. Trace the model's logic to a particular contextual vector and syntactic anomaly. Development of protocols for algorithmic explainability will thus be required governance functions. This demands the construction of localisable, interpretable machine learning architectures in conjunction with generative models, and all AI-driven conclusions must be accompanied by a transparent, rule-based reason that meets regulatory requirements.

5. Architecting the Future: Continuous Monitoring and Immutable Evidence

LLMs will be applied to conduct ongoing, real-time compliance monitoring of the audit work. With the continuous generation of unstructured data, the previous system of annual audits can no longer detect problems in time. The application of AI can set up an uninterrupted data pipeline for internal and external audit functions to automatically classify and assess documents at the time of creation. Thus, the audit period will no longer be after an incident but will be preventative and forward-looking.

The AI system that will continuously monitor the system must also be subject to continuous evaluation. Alterations in the business environment, economy and regulations are resulting in model concept drift. Therefore, the LLM will incorrectly interpret the altered contractual provisions in a subsequent period following its initial identification of lease obligations. Continuously monitor the drift in performance and adjust the extraction parameters of the model accordingly; therefore, a joint team of accountants and data scientists will be established.

In addition, the output of LLMs needs to be grounded in reality and meet the criteria for sufficient and relevant audit evidence. A model that reports a discrepancy should also provide precise semantic references to the specific section of the unstructured source document where that conclusion can be found. Therefore, the human auditor will be responsible for the final interpretation of the algorithm and can promptly compare it with the original text.

Finally, according to professional standards, the generation of immutable audit evidence for the operation of AI is required. Audit logs need to record the exact prompts that were used, the retrieval steps taken, the guardrails that were triggered, and any administrative parameter modifications during the session. The entire logging system will help us recreate the technological process that was used in the audit and provide external regulators with the necessary transparency to verify the AI-assisted audit method.

An AI-assisted continuous monitoring system will be introduced in the future, requiring new skills and a changed organisation for auditors. Traditionally, junior auditors manually perform vouching and only partners conduct an overall check; now, large language models can process a large number of documents at a high speed. The role of the junior auditor has changed; now they are prompt engineers and algorithms-output validators who need a solid foundation in data science concepts, statistical probability and digital forensics. Adjust the training content to cultivate a multi-disciplinary talent group. At the same time, the organisation's structure should also include specialised technology risk professionals to inspect the algorithmic structure and ensure that the code of the continuous-monitoring platform is secure, free from bias, and meets new professional requirements.

6. Conclusion

Exploration of large language models in the context of unstructured audit data shows that the process of financial supervision is also being reshaped. Expand the scope of analysis from organised numerical ledgers to cover a large amount of corporate narrative data, and then use LLMs to detect irregularities and automate complex compliance verification procedures to build an all-round risk assessment system for auditors. Natural language processing can be applied to address the problem of manual examination and gain more in-depth knowledge of the organisation through continuous observation.

However, this new technology has many problems ethically, technically and legally. Given the problems of hallucination, bias and privacy risks in generative models, strict governance should be introduced. The reliability of the audit process still relies on the professional scepticism of the human auditor; thus, these models should be used as advanced interpretation instruments rather than independent decision-makers. Strictly follow the rules for citing and immutably record operational logs to ensure the reliability of the audit evidence.

Finally, the application of AI in auditing should be done cautiously and judiciously. Build a particular model for the company and hire multi-disciplinary talent with both accounting knowledge and data science skills. With the continuous expansion of the volume of unstructured data in the digital economy, Large Language Models will gradually shift from being an advantage to indispensable tools in the practice of auditing in the future.

References

[1] Amoatey, D. K. (2026). The application of natural language processing towards auditing of unstructured data: A design science approach [Master’s thesis]. East Tennessee State University.

[2] Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. Intelligent Systems in Accounting, Finance and Management, 23(3), 157–214. https://doi.org/10.1002/isaf.1386

[3] Kim, A., Muhn, M., Nikolaev, V. V., & Tan, I. (2024). Large language models and financial reporting oversight [Conference presentation]. PCAOB/TAR Registered Reports Conference, PCAOB.

[4] Protiviti. (2023). Application of natural language processing in the context of internal audit. Protiviti.

[5] European Securities and Markets Authority, Institut Louis Bachelier, & The Alan Turing Institute. (2025). Leveraging large language models in finance: Pathways to responsible adoption. ESMA.

[6] Föhr, T. L., Schreyer, M., Moffitt, K. C., & Marten, K.-U. (2026). Deep learning meets risk-based auditing: A holistic framework for leveraging foundation and task-specific models in audit procedures. International Journal of Accounting Information Systems, 57, 100758. https://doi.org/10.1016/j.accinf.2026.100758

[7] Li, H., Gao, H., Wu, C., & Vasarhelyi, M. A. (2025). Extracting financial data from unstructured sources: Leveraging large language models. Journal of Information Systems, 39(1), 135–156. https://doi.org/10.2308/isys-2023-0143

[8] Fotoh, L. E., & Mugwira, T. (2025). Exploring large language models in external audits: Implications and ethical considerations. International Journal of Accounting Information Systems, 56, 100748. https://doi.org/10.1016/j.accinf.2025.100748

[9] Institute of Internal Auditors. (2024). Artificial intelligence auditing framework. The Institute of Internal Auditors.

[10] Dong, M. M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. International Journal of Accounting Information Systems, 55, 100715. https://doi.org/10.1016/j.accinf.2024.100715

[11] Kokina, J., Blanchette, S., Davenport, T. H., & Pachamanova, D. (2025). Challenges and opportunities for artificial intelligence in auditing: Evidence from the field. International Journal of Accounting Information Systems, 56, 100734. https://doi.org/10.1016/j.accinf.2025.100734