https://doi.org/10.58245/ipsi.tir.2603.05

Generative AI in Audit Oversight: A Checklist-Based Model Evaluation

Kresović, Dejana; Drulović, Sofija; Đoković; Nebojša, and Ćetković, Miloš

Abstract: ***This paper examines the use of generative artificial intelligence (GenAI) in audit oversight, with a focus on its role in the preliminary review of the independent auditor's report and accompanying financial statements. Existing audit supervision procedures rely on expert review, which is crucial but also time-consuming when it is necessary to examine large bodies of documents. The proposed solution is a framework supported by GenAI models, based on checklists, for identifying potential formal, structural and substantive problems in reporting that may require further expert review. The empirical analysis is based on 217 sets of documents of public companies in Serbia for the year 2024. ChatGPT and Gemini 3.1 Pro evaluated the same documents using a structured checklist and a risk scale of 0 to 3, while the expert assessment of audit supervision served as a benchmark. Model performance was evaluated using summary risk scores, mean absolute error, bias, Cohen's kappa coefficient, precision, responsiveness, F1-measure, accuracy, and ROC-AUC metrics. The results show that ChatGPT followed a more sensitive review pattern, with higher response and stronger adjusted agreement with the expert benchmark, while Gemini gave more conservative outputs, with higher precision but lower response. Both models performed better in formal and standardized checklist categories than in areas requiring professional judgment, such as key audit issues, business continuity considerations, and internal compliance red flags. The findings suggest that GenAI can support audit oversight by improving the consistency and prioritization of preliminary document review, but only when integrated into a structured framework with expert validation.***

Index Terms: ***audit oversight, ChatGPT, expert validation, Gemini, generative AI***

1. INTRODUCTION

The rapid development of generative artificial intelligence (GenAI) has opened up new opportunities for automation and support in professional document analysis [1]–[5]. In accounting and auditing, recent studies have explored the potential of large language models and AI-supported systems in audit documentation, analytical procedures, decision support, and professional work-flows [6]–[10]. In public administration, regulation and supervision, GenAI tools can contribute to faster, more consistent and more traceable preliminary analytical procedures. However, their use in specialized professional contexts requires caution, especially where model outputs may affect regulatory attention, public trust, or professional judgment [11]–[14].

Audit supervision is one such area. Independent auditor's reports are standardized professional documents, but they also include complex elements that depend on professional judgment [15]–[17], such as modified opinions, key audit matters, disclosures related to business continuity, emphasis of matter paragraphs, and internal consistency between the auditor's report and the accompanying financial statements [17]. These features make audit reports a useful environment for testing GenAI models, as some elements are formal and rule-based, while others require professional interpretation.

This paper does not address whether GenAI can replace auditors or audit oversight bodies. It can't. Rather, it examines whether GenAI models can support the preliminary review of audit reports by identifying potential formal, structural, and substantive issues that may require further expertise. This distinction is important because the value of GenAI in surveillance is not in autonomous decision-making, but in supporting the identification, documentation and prioritization of risks.

This paper evaluates the performance of two GenAI models, ChatGPT and Gemini 3.1 Pro, in the task of reviewing audit reports based on a checklist. Audit reports are used as real-world test documents, while expert assessment of audit oversight serves as a benchmark. The goal is not to assess the quality of individual audit firms, audit engagements or audit opinions. In-stead, the study examines how two GenAI mod-els perform in identifying potential issues in audit reporting within the same structured framework.

The contribution of this work is threefold. First, a framework for audit supervision supported by GenAI models, based on a structured checklist review, is proposed. Second, ChatGPT and Gemini are empirically compared on the corpus of audit reports and accompanying financial reports of public companies in Serbia. Third, model performance is evaluated using both concordance measures and binary problem detection metrics, distinguishing between sensitivity, con-

Manuscript received May 20, 2026

servatism, and concordance with expert judgment.

The rest of the paper is organized as follows. Section 2 presents the theoretical background and proposed framework. Section 3 explains the materials and methods. Section 4 presents and discusses the results. Section 5 concludes the paper and indicates the contribution and limitations.

2. Background And Proposed Framework

Independent auditor's reports are standardized professional documents used to communicate the auditor's opinion and selected reporting matters relevant to users of financial statements. Their structure and content are primarily determined by international audit reporting standards, including the standard that regulates the formation of opinions and reporting on financial statements [15]. In addition to formal reporting elements, audit reports may also include areas that require a higher level of professional judgment, such as key audit issues and business continuity considerations, which depend on entity-specific interpretation [16], [17]. This combination of standardized structure and judgment-based content makes audit reports suitable real-world documents for evaluating the performance of GenAI models in checklist-based review tasks.

Traditional audit oversight relies on peer review, professional standards, inspection procedures and regulatory assessment. At the same time, recent literature on AI-supported auditing emphasizes that intelligent tools can support but not replace professional judgment [8], [12], [13]. These mechanisms are necessary, but they also require a lot of time and resources. In digital public administration, supervisory bodies increasingly need tools that can support the review of documents and help identify cases that may require more detailed attention from experts. GenAI models can contribute to this process when applied within a structured and controlled framework.

Existing approaches to audit oversight and audit supported by artificial intelligence can be grouped into four broad categories: manual expert inspection supported by standardized checklists, document review based on rules or templates, traditional audit analytics and audit management systems, and new tools supported by artificial intelligence for audit documentation, explanations and decision support. These approaches are useful, but either still rely heavily on manual expert review, or focus mainly on audit execution rather than audit oversight, or provide limited evidence of alignment between GenAI output and an independent expert benchmark. This study addresses that shortcoming by testing a structured review framework supported by GenAI models on real audit reports and evaluating the model's output against audit oversight expert judgment.

The proposed framework treats GenAI as an assistive tool rather than an autonomous evaluator, which is consistent with previous literature that emphasizes explainability (explainable AI), human supervision, and risk-aware use of AI in auditing [12]–[14]. Its purpose is to support a preliminary review of audit reports using a predefined checklist. The framework consists of five steps: document collection, checklist decomposition, assessment using GenAI models, expert benchmark assessment, and statistical validation.

Figure 1 shows the workflow used in this study. Audit reports and financial statements are first processed through a structured checklist and then independently evaluated by ChatGPT and Gemini. Their risk assessments and findings at the level of individual items are compared with the expert judgment of the audit supervisor and statistically validated. This means that GenAI models can only be used as screening tools when their outputs are compared to expert judgment.

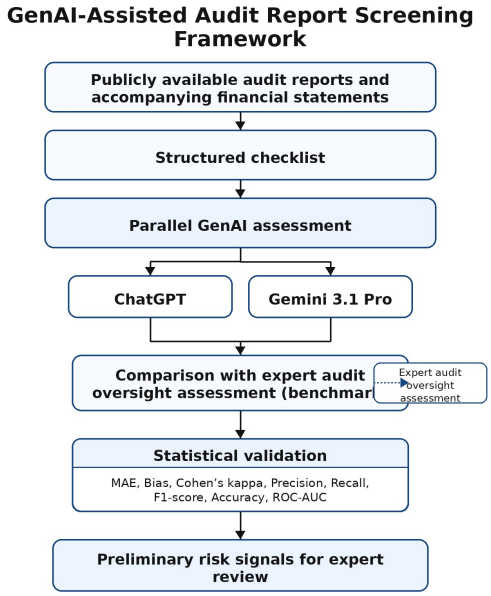

Figure 1. GenAI-Assisted Audit Report Screening Framework.

The framework is designed to be reproducible, transparent and suitable for empirical testing. Both GenAI models receive the same documents, the same checklist and the same rating scale. Their outputs are then compared to expert judgments, allowing the study to examine not only whether the models are detecting potential problems, but also whether their

judgments are aligned with professional supervisory judgment.

In this paper, the focus is on model performance. Audit reports are used as a real-world corpus to test the ChatGPT and Gemini models. The study does not rank audit firms, nor does it make final conclusions about the correctness of audit opinions. Instead, she examines whether GenAI models can support decision-making in the public sector by helping watchdogs identify reports that may require further peer review.

3. MATERIALS AND METHODS

3.1 Data and Audit Report Corpus

The initial research framework included 237 public companies in the Republic of Serbia. The empirical analysis was based on documents that are publicly available through the Agency for Business Registers (APR) [18]. During data collection, complete documentation was not available for 20 companies. These cases were therefore excluded, as they could not provide comparable textual inputs for assessment based on GenAI models.

The final empirical corpus consisted of 217 sets of documents. Each set of documents included one independent auditor's report and accompanying financial statements for the year 2024. The financial statement material included seven standard reporting components for each company, resulting in 1,519 financial statement components, in addition to 217 audit reports. A total of 1,736 documents or document components were used for document review based on GenAI models.

The audit reports were used as real-world test documents to evaluate the performance of the GenAI model. The objective was not to assess the quality of individual audit firms or audit opinions, but to examine how ChatGPT and Gemini identify issues in audit reporting based on the checklist when their outputs are compared to audit oversight expert judgment.

Table 1. Sample and assessment design

Element	Description
Research frame and exclusions	237 public companies; 20 excluded due to incomplete publicly available documentation on APR.
Final analytical sample	217 document sets.
Documents analyzed	Independent auditor's reports and accompanying 2024 financial statements.
Document components	217 audit reports; 1,519 financial statement components; 1,736 total components.
Models and benchmark	ChatGPT (GPT-5.5 Thinking) and Gemini 3.1 Pro, compared against expert audit oversight assessment.
Scoring and purpose	0-3 risk scale; evaluation of GenAI model performance in audit report screening.

3.2 Checklist and Scoring Operationalization

A structured checklist was developed to operationalize the preliminary review of the independent auditor's report and accompanying financial statements. The checklist followed the logic of international audit reporting standards and audit oversight practices, particularly ISA 700, ISA 701 and ISA 570 [15]–[17]. Its purpose was not to determine whether the audit opinion was correct or whether the audit work performed was sufficient, but to identify potential reporting problems that may require further expert review.

The checklist was organized into six categories: formal completeness of the audit report; structural coherence and clarity; consistency of audit opinion; key audit issues; business continuity and highlighting issues; and internal compliance, readability and red flags (risk indicators). Each category included several operational items that could be evaluated based on the audit report and accompanying financial statements.

Table 2. Checklist categories and operational focus

Category	Operational focus and assessment type
A. Formal completeness	Required sections and auditor/date identification; mostly formal.
B. Structure and clarity	Section order, wording clarity, and separation of opinion and basis; limited interpretation.
C. Opinion consistency	Alignment between opinion and basis, including modification wording; mixed formal/judgment-based.
D. Key audit matters	Relevance, entity-specific description, and explanation of auditor response; judgment-based.
E. Going concern/emphasis	Relevant disclosures, consistency with financial statement indicators, and material uncertainty clarity; judgment-based.
F. Internal consistency/red flags	Contradictions, missing explanations, unusual wording, and signals requiring expert review; interpretive.

Each item on the checklist was rated using a four-point ordinal scale of risk. A score of 0 indicated that no problem was identified. A score of 1 referred to a minor problem, such as unclear wording or limited formal inconsistency. A score of 2 indicated a moderate problem requiring further expert review. A score of 3 indicated a significant problem or potential regulatory issue. The same rating scale was applied by ChatGPT, Gemini and an expert rater.

The checklist was deliberately designed to include both standardized and judgment-based elements. Standardized elements are usually clearly visible in the audit report and can be assessed by reviewing the structure and wording of the document. Judgment-based elements require interpretation of the relationship between the auditor's report, financial statement disclosures and potential risk indicators. This design made it possible to examine whether GenAI models show different performance in formal review tasks compared to areas requiring professional judgment.

3.3 Expert Benchmark Assessment

The expert assessment of audit supervision served as a benchmark for evaluating the output of the GenAI model. The expert evaluator was a professional in the field of auditing and financial reporting with practical experience in reviewing independent auditors' reports, assessing compliance with audit reporting requirements and interpreting financial statements in the context of reporting by public interest companies or publicly available company reports.

The expert assessment was conducted independently of the assessment based on the GenAI models. The expert used the same checklist categories and the same rating scale from 0 to 3 as the two GenAI models. The purpose of this assessment was to establish a professional benchmark for preliminary problem detection, not to make final regulatory conclusions or re-evaluate the quality of the audit engagement itself.

To reduce the risk of bias, the expert rater did not use the outputs generated by ChatGPT or Gemini when assigning the reference scores. Therefore, the expert assessment was treated as an independent reference classification. GenAI outputs were compared to this benchmark only after all model evaluations were completed.

When the interpretation of a particular item on the checklist required professional judgment, the expert evaluator took into account both the audit report and the accompanying financial statements. Special attention is paid to modified opinions, key audit matters, disclosures related to business continuity, paragraphs that highlight a particular issue, inconsistencies between the auditor's report and information from the financial statements, as well as other red flags that may require further attention from supervisory bodies.

The quality control procedure was applied to a selected sub-sample of the evaluated document sets. A second auditor independently reviewed part of the sample to check the consistency of the expert assessment. All disagreements were discussed and resolved through professional judgment, with the original reference scores maintained or adjusted where warranted. This procedure was used to increase the reliability of the expert reference value and reduce the risk that the evaluation of the model depends solely on the interpretation of a single evaluator.

3.4 GenAI Assessment Protocol and Reproducibility Measures

An evaluation based on GenAI models was conducted using two general-purpose generative AI systems: ChatGPT, based on OpenAI's GPT-5.5 Thinking model, and Gemini 3.1 Pro, accessed through Google's Gemini environment. For better readability, the models are referred to as ChatGPT and Gemini in the rest of the paper.

Both models were evaluated under the same evaluation conditions. They were given the same document entries, the same checklist structure, the same rating scale, and the same task instructions. The models were instructed to act as audit oversight assistants and to perform only a preliminary review of the audit report and accompanying financial statements. They are specifically instructed not to determine whether the audit opinion is correct, not to assess the overall quality of the audit engagement, and not to make unsubstantiated regulatory conclusions.

The assessment protocol was designed to enhance comparability and reproducibility. For each set of documents, the relevant audit report and supporting material from the financial statements are made available to the model. The model was then asked to evaluate a set of documents using a predefined checklist and assign a score from 0 to 3 for each item on the checklist. For each non-zero score, the model was also asked to provide a brief rationale and to identify the element of the document that triggered the risk signal.

In order to reduce variability, the same prompt template was used for both models throughout the assessment. The query defined the role of the model, the scope of the task, the checklist categories, the rating scale, and the expected output format. The models were instructed to distinguish confirmed findings in documents from potential problems requiring expert review. They were also instructed to use careful language and to avoid conclusions that would normally require full audit inspection or regulatory authority.

The output of each model was recorded in a structured format, including checklist item scores, summary risk scores, and brief qualitative explanations. This structured output allowed for comparisons between the ChatGPT, Gemini, and expert benchmark models at both an aggregate and individual item level.

Since commercial GenAI systems may change over time, the model’s name, access environment, and evaluation period are documented for transparency and repeatability. This is especially important because the results of document analysis based on GenAI models may depend on query design, model version, document formatting, and interface-level settings that the researcher cannot always fully control.

Assessments were conducted during the same evaluation period using the ChatGPT and Gemini document upload web interfaces. No model output was used to modify the query during the main evaluation phase. The exact query template used in the study is listed in Appendix A.

3.5 Model Evaluation and Statistical Measures

The performance of the model was evaluated by comparing the output of the ChatGPT and Gemini models with an expert reference value (benchmark). The evaluation was conducted on two levels. First, the aggregated risk scores were used to compare the overall rating behavior of both models and the expert rater. Second, ratings at the level of individual checklist items were used to assess compliance and problem detection performance.

The mean absolute error (MAE) was calculated to measure the average distance between the model's rating and the expert's rating. Bias was used to determine whether the model generally overestimated or underestimated the risk relative to the expert reference value. A positive bias indicated that the model on average assigned a higher risk score than the expert, while a negative bias indicated a more conservative rating pattern.

Agreement was evaluated using item-level exact agreement and Cohen's kappa coefficient. Exact compliance measured the proportion of items on the checklist for which the model and the expert assigned the same rating. Cohen's kappa coefficient was applied because simple fit may exaggerate model performance when a large number of checklist items are classified as a problem-free situation [19].

For the purposes of binary problem detection analysis, the ordinal scores from the checklist were converted into two categories. A score of 0 was treated as a situation in which no problem was detected, while all scores greater than 0 were treated as a detected problem. Precision, recall, F1-score and accuracy were calculated using expert classification as a reference value. Accuracy measured the proportion of problems identified by the model that were also identified by the expert. The response measured the proportion of problems identified by the experts that the model was able to detect. The F1-measure was used as a balanced measure of precision and responsiveness, while accuracy measured the total proportion of correctly classified items.

The ROC-AUC metric was used to assess the ability of model scores to discriminate between lower and higher risk cases [20]. In the context of the preliminary review in audit supervision, the response (recall) was considered particularly important because the main purpose of the review is to identify cases that may require further expert analysis. Accuracy was also relevant, since a model that generates too many false positives can increase the workload of expert raters.

The evaluation therefore did not proceed from the assumption that assigning higher risk scores is automatically better. Instead, the analysis distinguished between sensitivity, conservatism, compliance with expert judgment, and practical utility for the first stage of review in audit oversight.

4. RESULTS AND DISCUSSION

4.1 Overall Risk-Scoring Patterns

The first step of the analysis compared the overall risk assessment behavior of the ChatGPT and Gemini models with an expert benchmark. The results show that the two models did not follow the same pattern. ChatGPT generally assigned higher risk scores, indicating a more sensitive screening approach. Gemini assigned lower scores and more often classified checklist items as non-problem situations, indicating a more conservative review pattern.

These findings should be interpreted in light of the study design itself. The models are not evaluated as autonomous decision-makers in the audit, but as structured preliminary review tools used under the same checklist, rating scale, and expert benchmark. Therefore, the results reflect their utility for risk-based document triage rather than their ability to make definitive conclusions in audit oversight.

Table 3. Overall model performance and agreement with the expert benchmark

Panel A. Overall risk-score distribution

Model/benchmark	Mean	Median	Min.	Max.
ChatGPT	0.428	0.45	0	0.88
Gemini	0.129	0	0	0.60
Expert assessment	0.218	0.22	0	0.73

Panel B. Agreement and error measures relative to expert be

Model/benchmark	MAE	Bias	Kappa	Interpretation
ChatGPT	0.204	0.204	0.532	More sensitive screening behavior
Gemini	0.123	-0.115	0.267	More conservative screening behavior
Expert assessment	-	-	-	Benchmark

Panel A. Overall risk-score distribution

Panel B. Agreement and error measures relative to expert benchmark

Table 3 summarizes the overall risk-score distribution and agreement measures. Panel A presents aggregate risk-score patterns, while Panel B reports model error, bias, and chance-adjusted agreement relative to the expert benchmark.

The results indicate that higher risk scores do not automatically imply better model performance. A model that assigns higher scores may be more sensitive, but may also be more likely to generate false positives. Conversely, a model that assigns lower scores may be more accurate in some cases, but may also miss problems identified by the expert. For this reason, evaluation of GenAI models in audit oversight should consider both sensitivity and accuracy.

The perceived pattern of evaluation is important for decision-making in the public sector. In a preliminary review in audit oversight, missing a relevant issue can be more problematic than flagging an item for further review. Therefore, a model with a higher recall may be more useful in the early stage of screening, while a more conservative model may be preferred only when the main priority is to reduce false alarms.

4.2 Agreement with Expert Assessment

The second step examined the level of agreement between each GenAI model and an expert reference value (benchmark). Exact compliance was calculated at the checklist item level, while Cohen's kappa coefficient was used to adjust for compliance that may have occurred by chance [19].

The results show that both models achieved some level of agreement with expert judgment, but the strength and nature of that agreement differed significantly. ChatGPT showed stronger adjusted concordance when Cohen's kappa coefficient was considered. In contrast, the Gemini model's compliance was partly affected by its frequent zero-risk classification. This suggests that Gemini can perform accurately in categories with a large number of items without problems, while being less reliable in detecting less obvious or judgmental problems.

In audit oversight, compliance with expert judgment is critical because the purpose of the GenAI model is not to make independent final decisions, but to support peer review. The results therefore support the use of GenAI models as an auxiliary review mechanism, provided that expert validation remains an integral part of the process.

4.3 Performance by Checklist Category

Model performance differed by checklist category. Both models scored better in standardized and formal categories than in areas requiring professional judgment. This pattern was expected, since formal elements are usually directly visible in an audit report, while judgmental elements require contextual interpretation.

Table 4. Performance by checklist category

Checklist category	ChatGPT agreement	Gemini agreement	Interpretation
A. Formal completeness	89.40%	97.10%	Strong standardized area
B. Structural compliance and clarity	80.00%	85.00%	Mostly standardized, moderate interpretive demand
C. Consistency of audit opinion	89.20%	91.70%	Relatively strong where wording is explicit
D. Key audit matters	66.20%	60.00%	Weakest area; requires professional judgment
E. Going concern and emphasis of matter	81.50%	78.30%	Context-dependent and judgment-based
F. Internal consistency and red flags	67.70%	66.70%	Requires interpretation and expert review

The best performances were observed in the categories related to formal completeness and report structure. These categories include elements such as title, opinion section, basis for opinion, management's responsibility, auditor's responsibility, date, and auditor identification. Since these elements were mostly explicitly stated in the report, both models identified them with relatively high consistency.

The weakest performance was recorded in key audit issues, business continuity considerations, as well as in internal compliance and red flag detection (risk indicators). These areas require more than just text recognition. They require an understanding of whether a matter is specific to a given entity, whether the auditor's explanation is sufficiently informative, whether the opinion is consistent with the basis for the opinion, and whether signals from the financial statements may require additional attention. The results therefore suggest that GenAI models are more reliable in standardized review tasks than in areas requiring professional skepticism and auditor judgment.

4.4 Issue-Detection Performance

In order to assess whether the models are capable of detecting problems in audit reporting, the ratings at the level of individual items were converted into binary classifications. A score of 0 was treated as a situation in which no problem was detected, while all scores greater than 0 were treated as a detected problem. Expert classification was used as a reference value (benchmark).

Table 5. Binary issue-detection performance

Model	Precision	Recall	F1-score	Accuracy	ROC-AUC
ChatGPT	0.551	0.874	0.676	0.830	0.746
Gemini	0.656	0.273	0.385	0.834	0.746

The results show that ChatGPT and Gemini differed significantly in their problem detection behavior. ChatGPT achieved a higher response (recall), which means that it detected a higher proportion of the problems identified by the expert. Gemini showed a more conservative detection pattern and flagged fewer problems. It achieved higher precision, but a significantly

lower response, indicating that it generated fewer false positives, but also missed a higher proportion of problems identified by the expert.

Interestingly, both models achieved the same ROC-AUC value, although their precision and response profiles differed significantly. This indicates that the models had a similar overall ability to distinguish between lower and higher risk cases, but differed in how they converted risk signals into problem detection. In practical terms, ChatGPT followed a more sensitive detection pattern, while Gemini took a more restrictive approach when flagging problems.

From an audit oversight perspective, responsiveness is particularly important in the preliminary review. The review tool should help identify cases that may require additional human analysis. If the model misses a large number of problems identified by the expert, its practical utility for supervision may be limited. Therefore, a model with a higher response may be more suitable for the first stage of screening, even if it generates more false positives. However, this advantage depends on the available expert review capacity, as high-sensitivity review may increase the number of cases requiring human verification.

The results also show that no model should be used without expert validation. GenAI models can support prioritization and documentation, but cannot determine whether the audit work performed was sufficient or whether the audit opinion was correct. Their role is limited to a structured preliminary examination.

4.5 Interpretation of ChatGPT and Gemini Behavior

Empirical results indicate that ChatGPT and Gemini should not be treated as interchangeable tools. ChatGPT acted as a high-sensitivity screening model. It tended to identify more potential problems, assign a higher risk score, and reveal a higher proportion of findings identified by the expert. This makes it more useful in situations where the main goal is to avoid missing relevant issues, provided there is sufficient expert review capacity to evaluate additional flagged cases.

Gemini behaved as a more conservative model. It produced fewer risk signals and assigned lower ratings more often. This may reduce the number of false positives, but may also increase the risk of underdetection (missing a problem). In audit oversight, this distinction is important because the cost of missing a relevant issue can be higher than the cost of sending a case for expert review.

The results suggest that GenAI tools can be useful for audit oversight only when integrated into a structured framework, which is consistent with previous research emphasizing human oversight, explainability, and ethical safeguards in AI-assisted auditing [12]–[14]. A checklist gives models a defined scope of work, a rating scale enables comparison, and an expert benchmark prevents undue reliance on AI-generated conclusions. This combination is crucial for the responsible use of GenAI models in public sector decision-making.

Overall, the proposed framework provides a practical way to test and document the performance of GenAI models in a specialized professional context. It does not replace human expertise, but it can improve the efficiency and consistency of preliminary document review. The findings also suggest that the choice of model should depend on the purpose of the review task itself: a more sensitive model may be preferred for early risk identification, while a more conservative model may be useful when reducing false positives is a priority and expert review capacity is limited.

5. CONCLUSION

From an operational point of view, the proposed framework offers a scalable and structured approach to the preliminary review of audit reports within a large corpus of documents. Its main advantage is the ability to consistently process standardized documents and identify cases that may require further expert review. However, the practical value of this framework depends on structured inquiries, pre-defined checklist criteria, transparent assessment, statistical validation and expert supervision.

The results indicate that ChatGPT and Gemini should not be treated as interchangeable tools. ChatGPT demonstrated higher sensitivity and significantly stronger recall, which makes it more suitable for first-stage screening tasks where the main goal is to avoid missing potentially relevant issues. However, this advantage depends on the available expert review capacity, since higher sensitivity may also increase the number of cases requiring human verification. Gemini gave more conservative outputs, with higher precision but lower response, which can be useful when reducing false positives is a priority.

Both models scored better in the formal and standardized categories of the checklist than in areas requiring professional judgment. Key audit issues, business continuity considerations and internal compliance red flags remained a more difficult task for both models. This suggests that GenAI can support audit oversight by improving the efficiency and consistency of the preliminary review, but it cannot replace audit oversight's expert judgment or professional judgment.

For decision-making in the public sector, the

proposed framework can improve the documentation, prioritization and consistency of preliminary reviews of audit reports. Its responsible use requires a human-in-the-loop approach, in which GenAI model outputs are treated as risk signals rather than final conclusions. In this sense, the framework is most useful as a decision support mechanism for identifying documents that deserve more detailed attention by experts.

The contribution of this work can be characterized through the creative methods of Combination and Interdisciplinary Transfer. The study combines generative artificial intelligence, audit supervision, expert validation and statistical evaluation of model performance within a single framework. She also applies document analysis based on GenAI models to the specialized context of audit report review and decision support in the public sector.

The study has several limitations. The analysis was carried out on audit reports and financial reports of public companies in Serbia for one reporting year, and the expert reference value was based on a professional assessment within the defined framework of the checklist. Results may therefore depend on the document corpus, checklist design, query structure, model version, and assessment context. Future research may extend this framework to other jurisdictions, additional GenAI models, multilingual audit reports, and other forms of regulatory documentation. Future studies should also examine whether optimized queries, structured inputs, multiple expert benchmarks, or hybrid review procedures (human-AI) can improve model performance.

Appendix A. Prompt Template Used for Genai Assessment

The following query template was used for both GenAI models during the main evaluation phase:

You act as an assistant in auditing supervision. Your task is to perform a preliminary review of the independent auditor's report and accompanying financial statements. You may not determine whether the audit opinion is correct, nor may you evaluate the overall quality of the audit engagement. Your role is limited to identifying potential issues in audit reporting that may require further expert review.

Use the following checklist categories: A. Formal completeness of the audit report; B. Structural coherence and clarity; C. Consistency of audit opinion; D. Key Audit Issues; E. Continuity of business and highlighting issues; F. Internal compliance, readability and red flags (risk indicators).

For each item on the checklist, assign a score as follows: 0 = no problem identified; 1 = minor problem; 2 = moderate problem that requires further expert review; 3 = significant problem or potential regulatory issue.

For each score greater than 0, please provide a brief explanation based solely on the documents submitted. Make a clear distinction between confirmed findings in the document and potential problems that require expert review. Do not draw unfounded regulatory conclusions.

References

[1] L. Banh and G. Strobel, “Generative artificial intelligence,” Electronic Markets, vol. 33, no. 1, article 63, 2023, doi: 10.1007/s12525-023-00680-1.

[2] J. Achiam et al., “GPT-4 Technical Report,” arXiv preprint, arXiv:2303.08774, 2023.

[3] Gemini Team et al., “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” arXiv preprint, arXiv:2403.05530, 2024.

[4] W. M. Lim et al., “Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators,” The International Journal of Management Education, vol. 21, no. 2, article 100790, 2023, doi: 10.1016/j.ijme.2023.100790.

[5] K. B. Ooi et al., “The potential of generative artificial intelligence across disciplines: Perspectives and future directions,” Journal of Computer Information Systems, vol. 65, no. 1, pp. 76–107, 2025, doi: 10.1080/08874417.2024.2302744.

[6] M. A. Vasarhelyi, K. C. Moffitt, T. Stewart, and D. Sunderland, “Large language models: An emerging technology in accounting,” Journal of Emerging Technologies in Accounting, vol. 20, no. 2, pp. 1–10, 2023, doi: 10.2308/JETA-2023-014.

[7] M. Eulerich, A. Sanatizadeh, H. Vakilzadeh, and D. A. Wood, “Is it all hype? ChatGPT’s performance and disruptive potential in the accounting and auditing industries,” Review of Accounting Studies, vol. 29, no. 3, pp. 2318–2349, 2024, doi: 10.1007/s11142-024-09833-9.

[8] H. Gu, M. Schreyer, K. Moffitt, and M. A. Vasarhelyi, “Artificial intelligence co-piloted auditing,” International Journal of Accounting Information Systems, vol. 54, article 100698, 2024, doi: 10.1016/j.accinf.2024.100698.

[9] J. Kokina, S. Blanchette, T. H. Davenport, and D. Pachamanova, “Challenges and opportunities for artificial intelligence in auditing: Evidence from the field,” International Journal of Accounting Information Systems, vol. 56, article 100734, 2025, doi: 10.1016/j.accinf.2025.100734.

[10] A. R. Otero and M. Agu, “Benefits and drawbacks of incorporating ChatGPT in financial audits,” Current Issues in Auditing, 2025, doi: 10.2308/CIIA-2024-015.

[11] A. Eisikovits, W. C. Johnson, and A. Markelevich, “Should accountants be afraid of AI? Risks and opportunities of incorporating artificial intelligence into accounting and auditing,” Accounting Horizons, 2024, doi: 10.2308/HORIZONS-2024-001.

[12] I. Munoko, H. L. Brown-Liburd, and M. A. Vasarhelyi, “The ethical implications of using artificial intelligence in auditing,” Journal of Business Ethics, vol. 167, no. 2, pp. 209–234, 2020, doi: 10.1007/s10551-019-04407-3.

[13] C. Zhong and S. Goel, “Transparent AI in auditing through explainable AI,” Current Issues in Auditing, vol. 18, no. 2, pp. A1–A14, 2024, doi: 10.2308/CIIA-2022-057.

[14] A. Birhane, R. Steed, V. Ojewale, B. Vecchione, and I. D. Raji, “AI auditing: The broken bus on the road to AI accountability,” in Proc. 2024 IEEE Conference on Secure and Trustworthy Machine Learning, 2024, pp. 612–643.

[15] International Auditing and Assurance Standards Board, International Standard on Auditing 700 (Revised),

Forming an Opinion and Reporting on Financial Statements. New York: IFAC, 2020.
[16] International Auditing and Assurance Standards Board, International Standard on Auditing 701, Communicating Key Audit Matters in the Independent Auditor's Report. New York: IFAC, 2015.
[17] International Auditing and Assurance Standards Board, International Standard on Auditing 570 (Revised), Going Concern. New York: IFAC, 2015.
[18] Serbian Business Registers Agency, Financial Statements Database. Belgrade: SBRA/APR, 2024.
[19] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.
[20] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

Dejana Kresović is a doctoral student at the Faculty of Organizational Sciences, University of Belgrade, and is employed at the Securities Commission of the Republic of Serbia in the field of public audit oversight and audit quality control. Her professional work includes the oversight of audit firms and licensed certified auditors, the application of international auditing standards, and the analysis of audit report quality. She is the author of several research papers and has presented at national and international scientific conferences. Her research interests include auditing, public oversight, financial reporting, and the application of generative artificial intelligence. E-mail: dejana.kresovic@sec.gov.rs, ORCID 0009-0008-9798-0733.

Sofija Drulović is a graduate of the Faculty of Organizational Sciences, University of Belgrade, and works as a Junior System Analyst at the Office for Information Technologies and eGovernment of the Republic of Serbia. Her professional focus includes information systems, business analysis, marketing, and strategic problem-solving. During her studies, she was actively involved in the Case Study Club and participated in local, national, and international case study competitions. Her experience includes marketing research, consulting projects, and the preparation of strategic recommendations for business clients. Her professional interests include digital transformation, e-government, business analysis, market research, and the application of artificial intelligence. E-mail: sofija.vukmirovic@ite.gov.rs, ORCID 0009-0003-7367-3230.

Nebojša Đoković holds a MSc in Economics and is an economic affairs adviser at the Association of Montenegrin Banks. He previously worked in banking supervision at the National Bank of Yugoslavia and the Central Bank of Montenegro, and later as an internal auditor at Hipotekarna banka. His professional experience includes banking regulation, financial supervision, internal audit, deposit protection, and European integration in the field of banking and financial services. He is the author of the book Banking of Montenegro and numerous professional articles. His interests include banking, financial services, economic policy, and European financial integration. E-mail: nebojsa.djokovic@mfa.gov.me, ORCID 0009-0002-9335-1530.

Miloš Cetković holds a degree in Management and a master's degree in Business Analytics from the Faculty of Organizational Sciences, University of Belgrade. His academic and research interests focus on retail development strategies, omnichannel business models, digital transformation, and the competitiveness of physical stores in the contemporary e-commerce environment. He has more than 25 years of professional experience in trade, retail, and international business. His career includes founding a company for the import and wholesale of computer equipment and developing a retail network of ten physical stores and four online sales channels in Serbia, Montenegro, and Bosnia and Herzegovina. E-mail: milos.cetkovic@yahoo.com, ORCID 0009-0000-9537-4651.