Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Human–artificial intelligence collaboration in professional judgment: large language model-assisted journal entry testing in financial audit

**László Péter Lakatos**

lakatos@uni-corvinus.hu

Corvinus University of Budapest

**Mária Laura Szabó**

Corvinus University of Budapest

**Article**

**Keywords:** human–AI collaboration, large language models, professional judgment, auditing, journal entry testing, accountability, fraud risk

**Posted Date:** November 25th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11486222/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

The delegation of professional judgment to artificial intelligence raises two questions that are especially pressing for professions built on documented judgment: which parts of expert work can be entrusted to algorithmic systems, and how can accountability be preserved when they are? This study examines these questions in external financial auditing, using journal entry testing (JET), a fraud-risk procedure that combines large-scale data screening with professional judgment, as a procedure-level case. A large language model (LLM) was instructed to screen a synthetic population of 1,000 journal entries against predefined risk indicators under a baseline prompt, a prompt to retain only items it was certain about, and a prompt to broaden the screen. An initial run was followed by six independent replication sessions and four sessions with paraphrased baseline prompts, with every flagged transaction exported for comparison. Repeating the identical baseline instruction produced near-identical selections (mean pairwise overlap 94.6%), and the fully specified weekend criterion returned the same count in every baseline, broadened and paraphrased session. Paraphrasing the instruction reduced overlap with the original selections to 87.1% on average, and paraphrased sessions overlapped with each other by 82.5%, whereas the certainty instruction produced unstable selections (51.9% overlap; 48 to 148 items retained). Where the instruction left a parameter open, the model resolved it sometimes consistently, as when it treated all pending and rejected entries as unapproved, and sometimes inconsistently, as with the jurisdiction of the holiday calendar or the window for period-end entries. The results suggest that in judgment-intensive screening, LLM output depends on how the request is framed and on interpretive choices that the instruction leaves to the model, with implications for documentation, review and accountability in AI-assisted professional work. The study does not assess fraud-detection accuracy; it offers a transparent, replicable design for locating where instruction-dependence enters an audit procedure.


# 1. Introduction

Artificial intelligence is rapidly entering areas of work where traditionally a significant professional judgment was unavoidable, with medicine, law and financial control among the most prominent examples (Susskind and Susskind, 2015; Dwivedi et al., 2021). This development raises a question of broad social-scientific significance: when an algorithmic system participates in expert decision-making, which elements of the task can be delegated safely, and how do a profession's accountability — documentation, review, and personal responsibility — absorb outputs whose content depends on how the system is instructed? External financial auditing offers a particularly instructive setting for this question, because it is a profession whose social function rests on documented, reviewable and attributable judgment. Therefore the strictly documented audit trail is unavoidable.

Journal entry testing (JET) is a central audit procedure in the assessment of fraud risk and management override of controls. It requires auditors to obtain, process, and evaluate large populations of unorganised accounting entries, often under substantial time pressure. In many engagements, the procedure combines rule-based filters, statistical or judgmental sample selection, spreadsheet work, and

manual review. The judgmental decisions (e.g. the choice of filtering criteria) are made on the basis of the experience and technical knowledge of more senior team members, while the execution of the testing is performed by assistants. This makes JET a suitable setting for studying whether artificial intelligence can improve audit workflow design without displacing professional judgment, as the filtering is still done by a human, while the execution might be delegated to an AI.

The practical challenge is not merely that JET is time-consuming. The deeper problem is that a sampling-oriented workflow may leave audit teams with a narrow view of the full transaction population, while a purely manual full-population review is usually infeasible. Monetary unit sampling and other sampling methods remain well-established and useful, but they are not primarily designed to identify all unusual behavioural patterns, especially when risk indicators are non-monetary, contextual, time-based, or user-specific.

Large language models (LLMs) create a new opportunity in this setting because they can process structured instructions, apply multiple risk criteria, generate explanations, and assist in documentation. The relevant question for audit research is therefore not whether an LLM can replace the auditor. It cannot and should not. The relevant question is whether an LLM can be embedded into a controlled, transparent, and human-supervised workflow that helps auditors prioritise which entries require further attention.

This study investigates that question through a controlled audit workflow experiment. A synthetic dataset of 1,000 journal entries was generated to reflect the type of information auditors typically receive during JET, including transaction identifiers, posting dates, account descriptions, monetary values, user identifiers, approval status, and approval timestamps. The LLM was then instructed to apply predefined risk indicators and to narrow the population into a smaller set of higher-priority entries for further manual review.

The study deliberately adopts a cautious contribution claim. It does not claim that the LLM improves fraud detection accuracy in a statistically conclusive sense. Such a claim would require independently validated labels, expert review, a confusion matrix, and paired statistical comparison against baseline methods. Its empirical claim is instead bounded to what a controlled design can show: how the output of an LLM responds to repetition, rewording and explicit changes in the selection threshold of its instructions, and how an LLM-assisted JET workflow can be structured, documented, and bounded so that it becomes suitable for further empirical testing in audit settings.

The paper makes three contributions. First, it provides replicated, transaction-level evidence on which parts of a concrete screening task an LLM reproduces consistently and which depend on how it is instructed, and it separates three sources of variation that are usually conflated: repetition of an identical instruction, rewording of the same task, and explicit changes in the selection threshold. Second, it shows that where an instruction leaves a parameter open, such as the jurisdiction of a holiday calendar or the window for period-end entries, the model fills the gap with definitional choices that can differ between sessions. This suggests that in AI-assisted screening part of the judgment that auditors would otherwise

exercise is exercised implicitly, through the wording of the instruction and the model's interpretation of it. Third, it draws a governance implication for accountable professions: if instructions and their interpretation shape the selection, the instructions and the criteria the model actually applied should become part of the documented and reviewable audit evidence, a point of direct relevance to the revised international standard on fraud (IAASB, 2025) and to audit documentation requirements (IAASB, 2009). The JET workflow and the replication design are a practical contribution that allows these claims to be tested on real or independently labelled data.

Beyond auditing, the paper speaks to the interdisciplinary literature on human–AI collaboration and the future of professional work (Jarrahi, 2018; Shrestha et al., 2019; Glikson and Woolley, 2020). The research question guiding the study is: how can an LLM-assisted workflow support risk-prioritised journal entry testing in a controlled audit setting, and what does the behaviour of such a workflow reveal about which elements of professional judgment can be delegated to generative AI?

## 2. Literature review and research positioning

# 2.1 AI and audit in the current literature

The integration of high-level automatisation into financial auditing has developed significantly over the past decades. Early research primarily focused on rule-based systems and expert models that attempted to replicate auditor judgment in structured environments (O'Leary and O'Keefe, 1997; Baldwin et al., 2006), or on IT applications that reduce monotonous manual work in audits. While these approaches marked an important first step, they were inherently limited in their ability to scale and to handle the complexity of unstructured and high-volume financial data.

More recent literature highlights the transformative role of AI in auditing, especially in terms of improving efficiency, effectiveness, and overall audit quality (Appelbaum et al., 2017; Kokina and Davenport, 2017; Vasarhelyi et al., 2015; Issa et al., 2016). AI-based tools can automate repetitive tasks, reduce the likelihood of human error, and enable the analysis of entire datasets instead of relying on traditional sampling techniques (Sebele, 2023; Yao, 2024). This shift reflects broader trends in accounting, where data analytics and digital technologies are becoming integral to professional practice (Agustí and Orta-Pérez, 2024; Alles, 2015).

A particularly important recent development is the increase in the use of large language models (LLMs), which extend AI capabilities beyond purely numerical analysis. Unlike earlier models, LLMs are capable of interpreting textual information, generating explanations, and supporting decision-making processes in a more contextual way (OpenAI, 2023; Vaswani et al., 2017). This makes them especially relevant for audit tasks, where professional judgment, documentation, and the interpretation of both structured and unstructured data play a central role (Bommarito and Katz, 2023; Keding, 2021; Dwivedi et al., 2021).

Fraud detection remains one of the most extensively studied applications of AI in auditing. Machine learning techniques have been widely used to identify unusual patterns and deviations in financial

transactions, including potential indicators of fraud or control override (Chandola et al., 2009; Ngai et al., 2011; West and Bhattacharya, 2016; Bolton and Hand, 2002). Deep learning models, in particular, have shown strong capabilities in detecting subtle and complex irregularities within large datasets (Schreyer et al., 2022; Pang et al., 2021), often outperforming traditional statistical approaches in terms of scalability and detection accuracy.

At the same time, the literature also highlights several challenges related to the adoption of AI in auditing. One of the main barriers is the limited technical expertise of auditors, which can hinder the practical implementation of advanced analytical tools (Barta, 2018; Sutton et al., 2016). As a result, there is an increasing need for solutions that are both technologically advanced and user-friendly, allowing auditors to apply them without extensive programming knowledge (Janvrin et al., 2008). Reviews of AI in the internal audit function similarly report low adoption and the absence of comprehensive implementation frameworks, and identify commitment, access, capability and skills development as preconditions for effective use (Wassie and Lakatos, 2024).

The impact of AI on the audit profession itself is another important topic of discussion. While AI has the potential to improve efficiency and reduce costs, it also raises concerns about job displacement and the changing nature of audit work (Brynjolfsson and McAfee, 2014; Susskind and Susskind, 2015; Frey and Osborne, 2017). However, the literature generally suggests that AI is more likely to transform the profession than to replace it entirely, with new roles emerging in areas such as data analytics, IT auditing, and AI governance (Barta, 2018; Raisch and Krakowski, 2021). Sociological accounts similarly describe workplace algorithms as a new contested terrain of control (Kellogg et al., 2020), the meaning of which professionals actively negotiate when their jurisdiction is challenged (Scarbrough et al., 2024).

In this context, the concept of human–AI collaboration has gained increasing attention. Rather than viewing AI as a substitute for auditors, many studies emphasise its role as a complementary tool that can support human decision-making (Jarrahi, 2018; Umar et al., 2021; Shrestha et al., 2019; Wilson and Daugherty, 2018). In such a model, AI handles data-intensive tasks, while auditors remain responsible for interpretation, professional judgment, and final conclusions. Experimental evidence shows that such collaboration is shaped by systematic behavioural tendencies: decision-makers may discount algorithmic advice after seeing it (Dietvorst et al., 2015) or, conversely, defer to algorithmic judgment more than to human judgment (Logg et al., 2019).

Journal Entry Testing (JET) represents a particularly relevant area for the application of AI. Traditional JET approaches rely heavily on sampling methods, such as monetary unit sampling (MUS), which inherently limit the scope of testing (Loughran, 2010; Lakis and Masiulevicius, 2017; National Audit Office, 2001). Although these methods are widely accepted, they may fail to capture certain types of irregularities, especially those that are systematic or of smaller magnitude. AI-based approaches, by contrast, allow for full-population analysis and risk-based prioritisation, thereby increasing the likelihood of detecting relevant anomalies (Schreyer et al., 2022; Appelbaum et al., 2017; Issa et al., 2016).

Another frequently discussed limitation is the lack of transparency in advanced AI models, often referred to as the “black box” problem. Many models, particularly deep learning systems and LLMs, provide limited insight into how their outputs are generated (Guidotti et al., 2018; Rudin, 2019; Lipton, 2018). This raises important concerns regarding trust, explainability, and regulatory compliance in audit settings. In high-stakes professional settings, such opacity also complicates the assignment of responsibility for algorithm-informed decisions (Busuioc, 2021; Lo Piano, 2020).

## 2.2 AI, audit analytics and the evolution of audit evidence

Artificial intelligence has been discussed in auditing for several years, beginning with rule-based expert systems and later expanding into machine learning, deep learning, anomaly detection, and audit data analytics *(*Gu et al., 2024). Prior research has emphasised the potential of advanced analytical tools to process larger datasets, identify unusual patterns, and support audit quality. In the audit context, however, technological capability is not sufficient. A tool must also be explainable enough to support review, sufficiently documented to form part of the audit file, and appropriately constrained so that professional judgment remains with the auditor.

LLMs differ from many earlier audit analytics tools because they combine analytical and language capabilities *(*Wang et al., 2025). They can interpret instructions, apply qualitative rules, explain why an item was flagged, and help create traceable documentation. These features make them attractive for audit workflows that involve both structured data and professional reasoning. At the same time, LLMs raise specific concerns: outputs may vary across runs, the internal reasoning process is not fully observable, and model performance may depend heavily on prompt design, data quality, and the audit context. Emerging audit evidence points in the same direction: LLMs can pass professional accounting examinations and support audit tasks (Eulerich et al., 2024), yet experienced auditors remain hesitant to rely on AI-based evidence for complex estimates (Commerford et al., 2022), and their reliance increases when they can provide input to the system (Commerford et al., 2024).

## 2.3 Journal entry testing as a procedure-level use case

JET is a particularly relevant use case because it sits at the intersection of data processing, fraud-risk assessment, and professional judgment. Common risk indicators include unusual posting times, weekend or holiday entries, high-value round-number transactions, unapproved entries, duplicate entries, threshold avoidance, end-of-period entries, and unusual user behaviour. Many of these indicators can be specified in structured prompts or rules, but their audit significance still requires human evaluation.

Traditional JET workflows often rely on statistical sampling, judgmental selection, or rule-based filtering. These approaches are useful, but each has limitations. Sampling may miss rare but important items. Purely judgmental selection can be inconsistent. Simple rules may create large numbers of false positives. An LLM-assisted workflow may help by combining multiple criteria, producing explanations, and narrowing the population into a risk-prioritised subset. The central audit question is whether such a workflow can improve prioritisation without creating unjustified reliance on the model. Such arrangements correspond to emerging human-in-the-loop work configurations, in which screening tasks

are shared between humans and algorithms while accountability remains with the professional (Grønsund and Aanestad, 2020). Interview evidence from auditors at manager level or above in Big Four firms suggests that big data analytics can strengthen journal entry testing and risk identification, but that over-reliance on analytics output threatens professional scepticism (Abu Al Rob et al., 2026). Such perception-based evidence leaves open how the algorithmic component itself behaves when the instructions given to it change, which is the focus of the present study.

## 2.4 Professional judgment, jurisdiction and accountability

Which elements of audit work can be delegated to an algorithm is not only a technical question. Sociological accounts of the professions locate professional authority in control over abstract knowledge and over the judgment that applies it to particular cases; tasks that can be routinised and specified as rules are those most exposed to jurisdictional competition (Abbott, 1988). From this perspective, the distinction between JET criteria that an instruction can specify completely and criteria that leave parameters open to interpretation is also a distinction between parts of audit work that can be codified and transferred to a machine and parts on which the profession's claim to expertise rests. Professionals actively negotiate this boundary when epistemic technologies enter their work (Kellogg et al., 2020; Scarbrough et al., 2024).

Auditing adds a second dimension. Its social function depends on verification that is documented, reviewable and attributable to an identifiable professional. Power's (1997) analysis of the audit society shows that the authority of audit rests less on direct observation than on procedures that render activities auditable and judgments defensible. Accountability, in turn, presupposes that an actor can be called to explain and justify conduct before a forum (Bovens, 2007). When an algorithm contributes to an audit judgment, the question becomes what must be documented for such justification to remain possible (Busuioc, 2021). If the content of an LLM's output depends on how it is instructed, the instructions themselves become part of the reasoning that has to be explained.

These questions are increasingly practical. ISA 240 (Revised), which applies to audits of financial statements for periods beginning on or after 15 December 2026, reinforces professional scepticism and the auditor's responses to the risk of fraud, including the risk of management override of controls that journal entry testing addresses (IAASB, 2025). ISA 230 requires audit documentation sufficient to enable an experienced auditor, having no previous connection with the audit, to understand the procedures performed and the significant professional judgments made (IAASB, 2009). Both standards are framed in technology-neutral terms and do not specifically address how the instructions given to a generative AI system should be documented when such a system takes part in selecting journal entries.

## 2.5 Research gap

Much of the existing literature discusses AI in auditing at a broad level. Fewer studies show how a specific audit procedure can be redesigned around AI while preserving auditability, documentation, and human oversight. There is also a need for more careful separation between feasibility evidence and effectiveness evidence. A workflow may reduce the number of items requiring review, but this alone does

not prove that it improves fraud detection or audit assurance. There is also little evidence on how the output of generative AI responds to the way professionals instruct it, although this determines where responsibility for the result lies.

This paper addresses that gap by presenting a procedure-level workflow for LLM-assisted JET and by explicitly limiting the conclusions to what the research design can support. The study therefore avoids claiming that LLMs replace sampling or auditor judgment. Instead, it develops a testable framework for risk prioritisation that can be subjected to stronger empirical validation in future work.

# 3. Methods

## 3.1 Research design

The study uses a controlled proof-of-concept design. The objective is to evaluate whether a structured LLM-assisted workflow can reduce and prioritise a journal-entry population for subsequent auditor review, and how its output responds to repetition of the same instruction, to rewording of the same task and to explicit changes in the selection threshold. The study comprises an initial run (11 May 2026) and a replication (23 October 2026). Generative AI is the object of the study and part of its workflow (dataset generation and screening; Section 6). Claude (Anthropic) was used to assist in revising the manuscript and in writing the analysis code; the authors designed and conducted the study, checked all content and remain responsible for it. The design is intentionally bounded: it examines workflow feasibility, risk prioritisation and prompt sensitivity, not conclusive fraud detection performance.

The workflow consists of five stages. First, a synthetic journal-entry population is prepared. Second, a set of predefined JET risk indicators is specified. Third, the LLM applies these indicators through structured prompts. Fourth, the model is instructed to narrow its first-round selection to the items it is certain about and, in a third round, to broaden it. Fifth, the LLM-assisted workflow is compared conceptually with a traditional sampling-oriented workflow at the level of population coverage, manual review burden, documentation needs, and implementation risks.

## 3.2 Dataset

The dataset contains 1,000 synthetic transactions designed to resemble journal-entry data used in external audit procedures. Each transaction includes a transaction identifier, timestamp, account information, debit or credit amount, journal-entry description, user identifier, approval status, and approval timestamp. The use of synthetic data avoids confidentiality concerns and enables a controlled demonstration of the proposed workflow.

The dataset includes risk patterns commonly considered in JET, including weekend and holiday postings (the generating model applied the England and Wales bank holiday calendar, although the generation prompt referred to the UK calendar), unapproved transactions, threshold-avoidance entries, end-of-period entries, large round-number transactions, or duplicate entries and suspicious user activity. Because the

dataset is synthetic and was not independently labelled by external experts, the study treats these patterns as embedded risk indicators rather than as confirmed fraud labels. The counts in Table 1 represent the embedded rule-based indicators in the synthetic dataset. The embedded indicators and their definitions were produced by the generating model, not specified by the authors; they are documented in the data dictionary supplied with the dataset (Supplementary Data 2; Supplementary Note 1). Four of them were re-derived deterministically from the data and matched the embedded flags row by row: weekend postings (Saturday or Sunday), holiday postings (2025 England and Wales bank holidays), unapproved transactions (approval status “Unapproved”) and end-of-period entries (postings on or after the 28th day of a month). The remaining definitions are less precise; for example, threshold avoidance refers to amounts narrowly below internal thresholds of EUR 10,000, 25,000, 50,000 or 100,000, and suspicious user activity to high-risk user behaviour such as postings by designated higher-risk users outside normal hours. Comparisons with the embedded indicators are therefore comparisons with the rule set of the generating model, not with an independent reference standard.

Table 1
– Flags and their count in the synthetic dataset

| Flag | Count |
|---|---|
| Weekend postings | 79 |
| UK bank holiday postings | 20 |
| Non-approved transactions | 55 |
| Threshold-avoidance entries | 71 |
| End-of-period entries | 191 |
| Large round-number transactions | 29 |
| Duplicate candidates | 40 |
| Suspicious descriptions | 93 |
| Suspicious user activity | 71 |
| Rows with at least one anomaly | 517 |
| Number of total anomalies in the dataset | 649 |

Table 1 describes the embedded characteristics of the synthetic dataset and should therefore be interpreted as the dataset-level reference structure used in the controlled experiment. The figures do not represent confirmed fraud cases, but predefined rule-based risk indicators intentionally included in the population. Some transactions contain more than one indicator, which explains why the total number of anomalies exceeds the number of rows with at least one anomaly.

## 3.3 LLM-assisted workflow and prompting protocol

The LLM was instructed to identify entries that matched predefined risk indicators. The baseline prompt (Section 6) listed the patterns to be flagged: weekend and holiday postings according to the UK holiday calendar, unapproved transactions, threshold-avoidance entries, end-of-period entries, and suspicious user activity, with descriptions, large round-number transactions and duplicate entries given as examples. It asked the model to flag the items, to comment on why each was suspicious and to tabulate the counts by category. It did not specify thresholds, time windows or other definitions.

The prompt design followed a human-in-the-loop logic. The LLM was not asked to conclude whether fraud existed. It was asked to prioritise entries that warranted further auditor attention, and potential human judgment, so that the auditor can examine and select items from the prioritised subset rather than from the entire population. This distinction is central to auditability: the model performs screening and documentation support, while the auditor remains responsible for evaluation, corroboration, and audit conclusions.

Three instructions were used, verbatim as documented in Section 6. The baseline prompt (B) asked the model to flag entries matching the listed indicators. The certainty prompt (S), issued in the same conversation, asked it to keep only items it was “100% sure” were suspicious, and the broadened prompt (X) asked it to extend the selection to every transaction that seemed “even a little bit suspicious”. The S and X prompts therefore deliberately change the selection threshold rather than merely the wording of the task. Because S and X were issued in the same conversation as B, as in the original protocol, their effect cannot be separated from the conversational context of the preceding baseline answer. To separate the effect of wording from that of the threshold, the replication added two paraphrased versions of the baseline prompt (P1 and P2), which request the same task with the same indicators in different words.

For the replication analysis, the risk criteria were classified according to how completely the instruction specifies them; the classification depends only on the wording of the instruction, not on the observed results. Weekend postings are fully specified: the posting date determines the result without further choices. For non-approved transactions and holiday postings, the instruction leaves a single parameter open, namely which approval statuses count as unapproved and which UK jurisdiction’s holiday calendar applies. The remaining criteria (threshold avoidance, end-of-period entries, suspicious descriptions, large round-number transactions, duplicates and suspicious user activity) require parameters that the instruction does not specify, such as threshold levels and proximity windows, the length of the period-end window, the meaning of suspicious wording or the rule for matching duplicates. A criterion counts as fully specified when the instruction, read literally, leaves no choice to the model. The classification compares different criteria; it does not vary the specification of the same criterion, which a future design should do.

## 3.4 Baseline comparison

The baseline comparison is a traditional JET workflow relying on spreadsheet preparation and sampling-oriented selection, particularly monetary unit sampling (MUS) or auditor judgment. The comparison is

not presented as a strict classifier-versus-classifier test, because the LLM workflow and the sampling workflow do not operate on identical decision logic. The LLM-assisted workflow screens the population using predefined risk indicators, whereas MUS increases the likelihood of selecting higher-value items. The comparison therefore focuses on workflow implications rather than on definitive detection superiority.

A future confirmatory study should include at least three baselines: a sampling-based workflow, a deterministic rule-based risk filter, and the LLM-assisted workflow. It should also use independently labelled data so that each method can be evaluated using precision, recall, false-negative rates, and paired statistical tests.

# 3.5 Evaluation approach

The evaluation in this paper is descriptive. It reports the number of entries flagged in each stage, the reduction in the candidate pool, and the practical implications for audit workflow design. It does not report accuracy, precision, recall, or false-negative rates, because those metrics require independent ground-truth labels that are not available in the current dataset.

For the replication, agreement between two sessions is measured at the transaction level as the Jaccard index, the number of transactions flagged in both sessions divided by the number flagged in either; the paper reports its mean and range across all pairs of sessions within a condition. Agreement with the embedded indicators is reported descriptively, as the share of the 517 items with at least one embedded indicator that a session did not flag. Differences in overlap between conditions are assessed with exact permutation tests, which make no distributional assumptions and remain valid for the small number of sessions. All tests are one-sided, with the alternative that overlap is lower in the comparison condition, at $\alpha = 0.05$; with three tests, the Bonferroni-adjusted threshold is 0.0167. For the baseline and paraphrased sessions, which are independent conversations, condition labels are reassigned across sessions. Because each certainty selection is derived from the baseline selection of the same conversation, the comparison between baseline and certainty selections uses a paired test that swaps labels only within conversations ($2^6 = 64$ configurations); its smallest attainable P value is 1/64.

To make the limitations transparent, the paper separates workload reduction from audit effectiveness. A reduction in flagged items shows only that the candidate pool has narrowed; it does not by itself show that audit effort decreases, because effort depends on how many items are subsequently tested and on the preparation, checking and documentation the workflow requires. Audit effectiveness, by contrast, would require evidence that relevant risk items are retained and that important items are not filtered out. That stronger claim is left for future validation.

# 3.6 Replication protocol

The replication was carried out on 23 October 2026 with GPT-5.6 Sol with high reasoning effort, selected once before the first session and not changed thereafter. Each session was run in a separate temporary conversation with memory, reference to earlier conversations and custom instructions disabled, so that

no session could draw on another. In each session the dataset without embedded indicators was uploaded. Six sessions repeated the three original prompts in their original order (B, S, X), and four sessions applied a paraphrased baseline prompt (P1 twice, P2 twice). After every prompt, a fixed export instruction asked the model to export the flagged transactions with one binary column per risk category and a short reason, without changing the selection (Section 6). This yielded 22 transaction-level files (Supplementary Data 3). In every baseline session, the dataset returned in the model-generated workbook matched the uploaded data row by row, and in each of the six sessions the certainty selection was a subset of the baseline selection, which in turn was a subset of the broadened selection. Because temporary conversations are not retained, the model label of individual sessions could not be verified afterwards; the model-generated workbooks, including the criteria the model reported applying, are provided as Supplementary Data 4.

## 4. Results

# 4.1 Initial run (11 May 2026)

The LLM-assisted workflow first flagged 564 of the 1,000 transactions as potentially relevant based on the predefined risk indicators. In a second stage, the model was instructed to prioritise the flagged items more narrowly. The resulting higher-priority subset contained 339 transactions, representing a 39.9% reduction relative to the first round and a 66.1% reduction relative to the full population. Section 4.2 shows that the size of this subset varied widely across replication sessions (48–148 items). This narrows the candidate pool, but it does not by itself demonstrate reduced audit effort or improved prioritisation, and it is not evidence of improved fraud detection.

In the initial run, criteria with no or a single open parameter returned identical counts under the baseline and the broadened prompt, while criteria with open parameters changed substantially depending on the requested depth of analysis. Table 2 reports the categories identified by the LLM during the prompting rounds. Differences between Table 1 and Table 2 therefore reflect model interpretation and prompt sensitivity rather than changes in the underlying dataset.

The criteria with no or a single open parameter (weekend postings, UK bank holiday postings and non-approved transactions) produced identical counts in the first and third round. For example, weekend postings remained at 79 items in the first and the third round, UK bank holiday postings remained at 35 in the first and third run, and non-approved transactions remained at 145. Within a single session, these categories were therefore stable; Section 4.2 shows how far this holds across sessions. Two of these stable counts nevertheless differ from the embedded indicators in Table 1, and both differences can be traced to definitional choices. The 145 non-approved transactions correspond exactly to the number of entries whose approval status is not “Approved” (55 unapproved, 64 pending and 26 rejected entries), whereas only the 55 unapproved entries were embedded as an indicator. The 35 bank holiday postings correspond exactly to the number of entries posted on a bank holiday in any part of the UK (England and Wales, Scotland or Northern Ireland), whereas the dataset was generated using the England and Wales

calendar, which yields 20 such entries. The stability of these criteria therefore reflects a consistent interpretation by the model rather than agreement with the embedded indicators, which shows that even criteria with a single open parameter involve interpretive decisions.

By contrast, categories with open parameters expanded substantially in the third round. Threshold-avoidance entries increased from 95 to 201, end-of-period entries from 179 to 394, suspicious descriptions from 96 to 337, and unusual user volume from 64 to 382. These results suggest that when the analysis is broadened to include items that are even slightly suspicious, the number of flagged entries increases considerably.

The high-confidence (second-round) prompt produced a different pattern. The model pruned the time-based and user-based categories (weekend postings by 84.8%, UK bank holiday postings by 82.9%, end-of-period entries by 67.6% and unusual user volume by 64.1%) while leaving the non-approved, threshold-avoidance, suspicious-description, round-number and duplicate counts unchanged. The asymmetry is therefore not simply that specified criteria are stable and underspecified criteria volatile: when asked for certainty, the model's notion of what is certainly suspicious did not follow the objectivity of the criterion. This further indicates that the output is shaped by the framing of the request rather than by the properties of the data alone.

Overall, the initial run indicates that criteria with no or a single open parameter provide stable outputs within a session when the scope of the prompt is broadened, while criteria with open parameters are highly sensitive to the depth and strictness of the analysis requested. Because this rests on a single session, it was examined in the replication. This distinction is important when interpreting AI-supported journal entry testing results. This pattern is consistent with technical evidence that LLM outputs are highly sensitive to seemingly minor features of prompt design (Sclar et al., 2024).

Table 2
– Baseline prompting, second-round prompting (narrowed-down sample) and third-round prompting (larger sample) results

| Category | First-round flagged items | Second-round flagged items | Reduction rate | Third-round flagged items | Expansion rate |
|---|---|---|---|---|---|
| Weekend postings | 79 | 12 | 84.81% | 79 | 0.00% |
| UK bank holiday postings | 35 | 6 | 82.86% | 35 | 0.00% |
| Non-approved transactions | 145 | 145 | 0.00% | 145 | 0.00% |
| Threshold-avoidance entries | 95 | 95 | 0.00% | 201 | 111.58% |
| End-of-period entries | 179 | 58 | 67.60% | 394 | 120.11% |
| Suspicious descriptions | 96 | 96 | 0.00% | 337 | 251.04% |
| Large round-number transactions | 35 | 35 | 0.00% | 56 | 60.00% |
| Potential duplicate entries | 40 | 40 | 0.00% | 59 | 47.50% |
| Unusual user volume | 64 | 23 | 64.06% | 382 | 496.88% |
| Number of flags | 768 | 510 | | 1688 | |
| Total number of flagged items | 564 | 339 | | 839 | |
| Total items | 1000 | 1000 | | 1000 | |

Table 2

reflects the model's classification and prioritisation behaviour under different prompting conditions. The first-round prompt asked the model to identify entries matching the specified JET risk indicators. The second-round prompt narrowed this output to entries the model treated as higher-confidence suspicious items, while the third-round prompt deliberately expanded the scope to include entries that appeared even slightly suspicious. The second-stage narrowing reduced the candidate pool to 339 items. This is a narrower pool, not a measured reduction in audit effort, and it is still too large for full manual review. The result should therefore be presented as an initial screening layer, not as a replacement for audit sampling or professional judgment.

| Result | First-round | Second-round | Third-round |
|---|---|---|---|
| Additional LLM flag outside embedded indicators | 86 | 61 | 322 |
| Missed embedded indicator | 39 | 239 | 0 |
| Agreement with embedded indicator structure | 875 | 700 | 678 |
| Total number of detected flagged items | 564 | 339 | 839 |
| Overall mismatch rate | 12.50% | 30.00% | 32.20% |
| Missed embedded-indicator rate | 7.54% | 46.23% | 0.00% |

Table 3 – Agreement between LLM outputs and embedded rule-based indicators

In the absence of independently validated fraud or risk labels, Table 3 should not be interpreted as a fraud-classification accuracy test. Instead, it reports the extent of agreement between the LLM outputs and the embedded rule-based indicators, that is, with the rule set of the model that generated the dataset. "Missed embedded indicator" refers to an item that contained at least one predefined synthetic risk indicator but was not retained by the LLM. "Additional LLM flag outside embedded indicators" refers to an item flagged by the LLM although it did not contain any of the predefined embedded indicators. Such additional flags may reflect broader model interpretation, over-inclusion, or limitations of the embedded rule set, rather than confirmed false positives. "Agreement with embedded indicator structure" refers to cases where the LLM output was consistent with the embedded-indicator structure. "Overall mismatch rate" captures all disagreements between the LLM output and the embedded indicators relative to the full population of 1,000 entries, while "missed embedded-indicator rate" captures the proportion of the 517 items with at least one embedded indicator that were not retained by the LLM.

The first-round screening appears to provide the most balanced output in this controlled setting. It retained 564 items for further auditor consideration and resulted in a missed embedded-indicator rate of 7.5%. This suggests a reasonable trade-off between coverage and review burden. By contrast, the third-round screening retained 839 items and achieved a 0% missed embedded-indicator rate in this dataset, but at the cost of a substantially larger review population. This result should not be interpreted as evidence that the third-round prompt would achieve complete coverage in other datasets. The second-round high-confidence subset retained only 339 items, but its missed embedded-indicator rate increased to 46.2%. Therefore, the second-round output appears more suitable as a high-priority triage layer than

as a standalone basis for audit testing, since almost half of the items with embedded risk indicators would remain outside the reviewed population.

Table 4
– Workflow-level comparison of LLM-assisted and traditional JET approaches (conceptual comparison; the sample size of about 25 items is an illustrative assumption)

| Dimension | LLM-assisted workflow | Traditional sampling-oriented workflow | Interpretation |
|---|---|---|---|
| Population coverage | Initial screening can be applied to the full 1,000-entry population | Selection is typically made from the full population using sampling or judgment | LLM support may increase screening breadth, but not necessarily assurance quality |
| Candidate pool and items tested | The narrowed pool (339 items in the initial run, 48–148 in the replication) is the candidate pool from which items are selected for testing (approximately 25 items, an illustrative assumption) | Approximately 25 selected items in the illustrative manual workflow | LLM screening narrows the candidate pool but does not reduce the number of items tested; effort savings and prioritisation quality were not measured |
| Decision logic | Risk-indicator prioritisation | Monetary or judgmental selection | The methods answer different audit questions and should not be treated as strict substitutes |
| Documentation | Requires prompt, model, criteria, output, and auditor review documentation | Requires sampling method, rationale, and working paper documentation | LLM use increases documentation complexity |
| Main risk | False comfort, prompt sensitivity, unexplained outputs, data confidentiality | Sampling risk, fatigue, limited coverage, subjective judgment | Both approaches require audit safeguards |

LLM support may complement, but should not replace, auditor judgment, documentation requirements, and conventional audit safeguards. Table 4 summarises the workflow-level implications of the LLM-assisted approach compared with a traditional sampling-oriented JET workflow. The purpose of the comparison is not to present the two methods as direct substitutes, but to clarify their different audit roles. The LLM-assisted workflow supports broader initial screening, while the traditional workflow provides a more established sampling-based testing approach.

# 4.2 Replication across sessions and prompt wordings (23 October 2026)

Table 5 summarises the replication at the transaction level. Repeating the identical baseline instruction in six independent sessions produced between 455 and 503 flagged items (mean 469), with a mean pairwise overlap of 94.6% (range 88.5–99.3%); 453 transactions were flagged in every session. The baseline selection is therefore largely reproducible across sessions, although it is smaller than in the initial run (564 items).

The certainty instruction behaved very differently. The six certainty selections ranged from 48 to 148 items (mean 96) and overlapped by only 51.9% on average (range 26.1–95.8%); of the 218 transactions retained in at least one session, only 48 were retained in all six. The broadened instruction was more stable (mean overlap 83.8%) but expanded the selection to between 543 and 681 items. In one broadened session the model added criteria that the instruction did not contain, chiefly entries posted through the manual journal channel (68 items) and postings outside normal hours; 113 of that session's 680 items fall under none of the nine categories. Excluding them leaves the mean overlap of the broadened selections practically unchanged (83.4% instead of 83.8%). Paraphrasing the baseline instruction changed the selection moderately: the four paraphrased sessions flagged between 429 and 569 items, their overlap with the six original baseline sessions averaged 87.1% (range 80.0–97.9%), and two sessions with the same paraphrase overlapped by 78.9% and 83.5%, less than repeated sessions with the original wording. Exact permutation tests over the ten baseline and paraphrased sessions confirm that overlap between paraphrased and original sessions is lower than overlap among the original sessions ($P = 0.010$, 210 permutations), as is overlap among the paraphrased sessions ($P = 0.010$); the lower stability of the certainty selections relative to the baseline selections is also significant in a paired test within conversations ($P = 0.016$, the smallest value attainable with six pairs). All three P values remain below the Bonferroni-adjusted threshold.

Within this design, the three conditions therefore differ in how much the selections vary. Repetition of an identical instruction produced the smallest differences, rewording of the same task produced moderate differences, and an instruction that appeals to the model's certainty produced the largest and least predictable differences. With two sessions per paraphrase, the size of the rewording effect is nevertheless estimated imprecisely, and because the certainty and broadened prompts always followed the baseline prompt in the same conversation, these comparisons describe this particular prompting sequence; they do not isolate the threshold instruction from conversational history and order. The changes between rounds in the initial run (Table 2) should therefore be read primarily as responses to explicit changes in the selection threshold, which exceed the effect of rewording the task, rather than as effects of framing alone.

Table 5
– Replication: transaction-level stability by prompt condition. Pairwise overlap is the Jaccard index between two sessions of the same condition.

| Condition | Sessions | Items flagged, mean (range) | Pairwise overlap, % mean (range) | Flagged in every session | Flagged in at least one session | Embedded-indicator items not flagged, % mean (range) |
|---|---|---|---|---|---|---|
| Baseline prompt (B) | 6 | 469 (455–503) | 94.6 (88.5–99.3) | 453 | 519 | 19.5 (13.3–21.7) |
| Certainty prompt (S) | 6 | 96 (48–148) | 51.9 (26.1–95.8) | 48 | 218 | 82.0 (73.5–90.7) |
| Broadened prompt (X) | 6 | 630 (543–681) | 83.8 (78.2–90.6) | 532 | 769 | 5.0 (3.3–8.3) |
| Paraphrased baseline (P1, P2) | 4 | 500 (429–569) | 82.5 (75.4–90.3) | 418 | 592 | 16.8 (6.6–28.6) |

Table 6 reports the counts per risk category. The fully specified weekend criterion returned 79 items, identical to the embedded indicator, in every baseline, broadened and paraphrased session. For the two criteria with a single open parameter, the model's choice was consistent in one case and not in the other. It treated every entry whose status was not "Approved" as non-approved in all 16 of these sessions, yielding 145 items rather than the 55 embedded unapproved entries, and stated this rule explicitly where it documented its criteria. For holiday postings, by contrast, it applied the England and Wales calendar in five baseline sessions (20 items) and the calendars of all UK nations in one baseline session and in all four paraphrased sessions (35 items).

Criteria with open parameters varied more, but not uniformly. End-of-period entries ranged from 96 to 183 across the baseline sessions and suspicious descriptions from 48 to 101, whereas threshold-avoidance entries (64–67) and duplicates (40 in every session) were resolved almost identically. Suspicious user activity illustrates an ambiguity in the instruction itself, which lists descriptions, round amounts and duplicates as examples of "suspicious user activity": in every baseline session the model used the term as an umbrella for these examples and created no separate user-level category, whereas under the broadened instruction it did create one (53 to 302 items). Under the certainty instruction, the model removed most weekend and holiday items and, in five of the six sessions, most non-approved items, but retained 39 or 40 threshold-avoidance entries in every session; its notion of certainty thus appears to track the apparent intent behind an entry rather than how completely the criterion was specified.

The definitional choices also account for most of the embedded-indicator items that the baseline sessions did not flag (19.5% of the 517 items on average). In every baseline session, 51 of the missed

items carried only the embedded suspicious-user-activity indicator, which the model did not operationalise as a separate category, and between 7 and 47 carried only the end-of-period indicator, reflecting a narrower period-end window than the embedded definition.

Table 6
– Replication: items flagged per risk category, range across sessions

| Risk category | Specification in the instruction | Embedded indicator | Baseline (6 sessions) | Certainty (6 sessions) | Broadened (6 sessions) | Paraphrased (4 sessions) |
|---|---|---|---|---|---|---|
| Weekend postings | fully specified | 79 | 79 | 0–8 | 79 | 79 |
| Non-approved transactions | one open parameter (statuses) | 55 | 145 | 7–87 | 145 | 145 |
| UK bank holiday postings | one open parameter (jurisdiction) | 20 | 20–35 | 0–1 | 20–35 | 35 |
| Threshold-avoidance entries | open parameters | 71 | 64–67 | 39–40 | 82–137 | 49–75 |
| End-of-period entries | open parameters | 191 | 96–183 | 2–21 | 183–300 | 96–183 |
| Suspicious descriptions | open parameters | 93 | 48–101 | 9–71 | 121–178 | 48–121 |
| Large round-number transactions | open parameters | 29 | 35–40 | 0–16 | 40 | 35–40 |
| Potential duplicate entries | open parameters | 40 | 40 | 0–40 | 40–44 | 2–40 |
| Suspicious user activity | ambiguous scope | 71 | 0 | 0 | 53–302 | 0–121 |

Compared with the initial run, the replication reproduced the contraction under the certainty prompt and the expansion under the broadened prompt, the identical weekend and non-approved counts under the baseline and broadened prompts, and the model's treatment of all non-approved statuses as unapproved. It did not reproduce the size of the certainty selection (339 items in the initial run against 48–148), the stability of five categories under the certainty prompt reported in Table 2, the separate category for unusual user volume (64 items in the initial run, none in any replication baseline session) or the UK-wide holiday calendar (used in one of six replication baseline sessions). The item-level selections underlying Tables 2 and 3 are not part of the supplementary material, so transaction-level overlap between the initial run and the replication is not reported. A further single session with a newer model

version, reported in Section 6, also reproduced the contraction and expansion (475, 120 and 777 items). The initial run is therefore best read as an exploratory pilot whose qualitative pattern, but not its exact figures, was replicated.

# 5. Discussion

The results suggest that an LLM can support JET as a screening layer. The workflow screened the full population and narrowed the candidate pool of entries requiring further attention, and with an identical instruction it did so reproducibly across sessions. This is useful because it may help audit teams focus on entries with clearer risk characteristics, especially when the initial population is too large for full manual inspection.

However, the findings should not be overstated. A reduction from 1,000 entries to 339 entries in the initial run does not prove that the LLM preserved all relevant risk items. Without independent labels, the study cannot determine whether important anomalies were missed. Nor can it establish whether the LLM outperformed MUS, rule-based filters, or expert judgment. The appropriate conclusion is that the workflow is feasible and potentially useful, not that it is empirically superior. Nor does a narrower candidate pool show that audit effort decreases, and the replication shows that the narrowest selection, produced by the certainty instruction, is also the least reproducible.

The study also highlights the importance of distinguishing between automation and audit judgment. The LLM can assist with screening, summarisation, and documentation, but final audit decisions require human evaluation. For example, a weekend posting may be suspicious in one entity and routine in another. An unapproved transaction may reflect a control deficiency, a timing issue, or a data extraction artifact. The model can flag such items, but it cannot determine their audit significance without contextual evidence. The human-factors literature underlines why this boundary matters: automation bias — the over-reliance on automated aids — is well documented and persists even when users are made accountable for their decisions (Parasuraman and Riley, 1997; Skitka et al., 1999, 2000; Goddard et al., 2012).

Beyond the audit setting, the asymmetries documented in Tables 2, 5 and 6 speak to a broader social-scientific debate about the nature of algorithmic output in judgment-intensive work. Repeating an instruction reproduced the selection, whereas rewording it, appealing to the model's certainty or leaving a parameter open changed what the model 'found'. This suggests that, in judgment-intensive screening, the result is co-produced by the human framing of the request and by interpretive choices the model makes where the request is incomplete. Professionals may nonetheless treat such output as an objective screening result, a tendency consistent with research on human trust in artificial intelligence (Glikson and Woolley, 2020) and with concerns about the interpretability of complex models in high-stakes decisions (Rudin, 2019; Guidotti et al., 2018). For the governance of AI-assisted professional work, the implication is that accountability frameworks must document not only the model and its output but also the instructions given to the model and the criteria it reports applying, because the instructions

are part of what produces the result. In terms of the jurisdictional view of the professions (Abbott, 1988), we propose, as a proposition for empirical testing rather than as a finding, that delegation does not simply transfer specifiable tasks to the machine while leaving judgment with the auditor: part of the judgment appears to migrate into the formulation of the instructions and into the model's interpretation of them, which may become a new site of professional expertise. The study observes this at the level of model output; how auditors themselves formulate instructions and justify the resulting selections remains to be examined. For audit as a practice of rendering work auditable (Power, 1997), the defensibility of an AI-assisted selection would then depend on whether this instruction-level judgment is itself made visible and reviewable.

For standard setters and oversight bodies, this has a concrete implication. ISA 230 requires documentation sufficient for an experienced auditor to understand the significant professional judgments made in reaching audit conclusions (IAASB, 2009). Where an LLM contributes to the selection of journal entries, the prompt and its revisions, the model version, the date of each run and the criteria the model reports having applied are part of how that selection was made, and they should therefore be retained in the audit file in the same way as the parameters of a conventional data-analytics filter. As ISA 240 (Revised) comes into effect (IAASB, 2025), guidance on documenting AI-assisted procedures would help audit firms and public oversight bodies distinguish workload reduction from evidential support.

For practice, the proposed workflow is most defensible as a complementary layer before sample selection or detailed testing. Audit teams could use the LLM-assisted output to identify risk clusters, inform judgmental selection, supplement statistical sampling, and document the rationale for further review. This use case is more credible than claiming that an LLM can replace JET sampling or eliminate the need for manual testing.

For research, the next step is clear. The workflow should be tested on a dataset with independently validated labels. At least two audit experts should label the entries, inter-rater agreement should be reported, disagreements should be reconciled, and the LLM should be evaluated against deterministic rules and sampling-based baselines. Only then can the paper move from feasibility evidence to effectiveness evidence.

## 6. Robustness, reproducibility and audit documentation requirements

This study is not designed to test model generalisability. It is designed to examine whether a structured LLM-mediated audit workflow can produce reproducible, reviewable, and auditor-controllable prioritisation outputs under controlled conditions. The following approach was therefore used.

The dataset was generated on 11 May 2026 with GPT-5.5 Thinking from the prompt reproduced in Supplementary Note 2. The conversation was then deleted (screenshots of it are retained by the authors)

and the session was reset, so that the model had no memory of the generation step. The initial JET run was carried out with the same model on the same day in a new session, using the baseline, certainty and broadened prompts reproduced verbatim in Supplementary Note 2. A further single session with GPT-5.6 Thinking in early autumn 2026 applied the same dataset and prompts (table below). The replication of 23 October 2026 used GPT-5.6 Sol with high reasoning effort under the protocol described in Section 3.6; the export instruction and the two paraphrased baseline prompts are also reproduced in Supplementary Note 2.

Reproducibility test on GPT-5.6 Thinking Model (early autumn 2026)

| Search approach | Unique transactions flagged | % of population |
|---|---|---|
| Initial analytical screening | 475 | 47.5% |
| Strict high-confidence review | 120 | 12.0% |
| Expanded broad screening | 777 | 77.7% |
| Total number of transactions: 1,000 | | |

The same dataset and prompts were used as in the first part of the article.

The newer model version flagged fewer items under all three prompts (475, 120 and 777, compared with 564, 339 and 839), but it reproduced the qualitative pattern: contraction under the high-confidence prompt and expansion under the broadened prompt. Run-to-run variability was examined in the replication described below.

# 7. Limitations and future research

This study has several limitations. First, it relies on a synthetic dataset. Synthetic data are useful for controlled experimentation and confidentiality protection, but they may not reflect the complexity, noise, and entity-specific context of real accounting systems. Second, the dataset does not contain independently validated fraud or risk labels. As a result, the study cannot calculate precision, recall, false-negative rates, or accuracy. This concern echoes evidence that even expert-generated ground-truth labels are contestable, which complicates the evaluation of AI tools in professional work (Lebovitz et al., 2021).

Third, the same general technological environment was involved in both dataset generation and evaluation. This creates a risk of construct overlap and potential circularity. Future studies should separate data generation, labelling, and model evaluation. Ideally, the evaluation dataset should be created or labelled independently of the LLM being tested.

Fourth, the current study does not address legal, confidentiality, and data protection issues in sufficient detail. In real audit engagements, client data should not be entered into public AI systems unless

confidentiality, access, retention, and training-use restrictions are contractually and technically controlled. Future work should examine secure deployment models, audit-firm governance, and documentation expectations.

Fifth, the replication comprises six sessions per prompt condition and four paraphrased sessions with a single model, and it was carried out five months after the initial run, so the two are not directly comparable in absolute terms. Because the sessions were run in temporary conversations, the model label of individual sessions could not be verified afterwards. Larger numbers of sessions, several model families and interfaces with fixed settings would allow the variation to be modelled formally.

Future research should test the workflow using real-world or independently labelled data and should compare LLM-assisted JET against sampling, deterministic rule-based filters, and machine-learning anomaly detection models. It should also investigate whether auditors using LLM-assisted workflows make better, faster, or more consistent decisions than auditors using traditional tools. Experimental studies with practitioners should examine how auditors formulate instructions for such systems and whether an independent reviewer can understand and justify the resulting selections, which is where the accountability questions raised here would be tested directly. A minimal test would ask an independent auditor to reconstruct and justify a session's selection from the exported prompt and the criteria the model reported applying, which corresponds to the "experienced auditor" criterion of ISA 230.

# 8. Conclusion

This paper presents a controlled audit workflow study of LLM-assisted journal entry testing. The results show that a structured LLM workflow can screen a full transaction population and reduce it to a smaller candidate pool for further auditor review. In the initial run, the certainty prompt narrowed the population from 1,000 entries to 339, but this figure did not replicate (48–148 items across six sessions), whereas the identical baseline instruction reproduced its selection with a mean overlap of 94.6%.

The study's primary contribution is to show, with replicated transaction-level evidence, which parts of a concrete audit screening task an LLM reproduces and which depend on how it is instructed. An identical instruction produced nearly identical selections; rewording the same task changed them moderately; an appeal to the model's certainty made them unstable; and where the instruction left a parameter open, the model resolved it with definitional choices that were consistent for some criteria and not for others. In judgment-intensive screening, LLM output therefore depends on the instructions it receives and on the model's interpretation of them. We propose, for empirical testing, that part of the judgment auditors would otherwise exercise shifts to the formulation of instructions and the specification of criteria. If so, LLM support would be least problematic where criteria can be fully specified and documented, while auditor judgment would remain indispensable for the specification of criteria, the evaluation of audit significance and the design of responses. It also implies that the instructions given to the model, and the criteria it reports applying, should be documented as part of the audit evidence. The documented,

human-supervised JET workflow and the replication design are practical contributions that make these claims testable on real or independently labelled data.

This contribution is relevant because audit practice increasingly requires methods that can handle large datasets while preserving professional accountability. LLM-assisted workflows may support this transition, but they must be implemented with appropriate safeguards, transparency, documentation, and human review. The paper therefore provides a foundation for future empirical research rather than a final claim of audit effectiveness.

## Declarations

# Competing interests

The authors declare no competing interests.

# Author Contribution

Authors: László Péter Lakatos[1] (corresponding author) and Mária Laura Szabó[2]. [1] Institute of Accounting and Law, Corvinus University of Budapest, Budapest, Hungary; [2] Institute of Accounting and Law, Corvinus University of Budapest, Budapest, Hungary.Author contributions: Conceptualization: L.P.L., M.L.S.; Methodology: L.P.L., M.L.S.; Investigation (dataset generation and initial run): M.L.S.; Investigation (replication sessions): L.P.L.; Formal analysis: L.P.L.; Writing – original draft: [M.L.S., L.P.L.]; Writing – review and editing: L.P.L., M.L.S.; Supervision: L.P.L.

# Data Availability

All data supporting the findings of this study are available within the paper and its Supplementary Information. Microsatellite primer sequences are provided in Supplementary_Data_0.xlsx, along with original reference describing the microsatellites used in this study.

## References


1. Abbott A (1988) The system of professions: an essay on the division of expert labor. University of Chicago Press, Chicago
2. Abu Al Rob MA, Mohd Nor MN, Salleh Z (2026) Utilizing big data analytics in the selection and testing of journal entries and its impact on professional skepticism. Humanit Social Sci Commun 13:553. https://doi.org/10.1057/s41599-026-06626-0
3. Agustí MA, Orta-Pérez M (2024) Big data and artificial intelligence in the fields of accounting and auditing: a bibliometric analysis. Span J Finance Acc. https://ssrn.com/abstract=4155537

4. Alles MG (2015) Drivers of the use and facilitators and obstacles of the evolution of big data by the audit profession. Acc Horizons 29(2):439–449. https://doi.org/10.2308/acch-51067
5. Appelbaum D, Kogan A, Vasarhelyi MA (2017) Big data and analytics in the modern audit engagement: research needs. Auditing: J Pract Theory 36(4):1–27. https://doi.org/10.2308/ajpt-51684
6. Baldwin AA, Brown CE, Trinkle BS (2006) Opportunities for artificial intelligence development in the accounting domain: the case for auditing. Intell Syst Acc Finance Manage 14(3):77–86. https://doi.org/10.1002/isaf.277
7. Barta G (2018) The increasing role of IT auditors in financial audit: risks and intelligent answers. Bus Manage Educ 16(1):81–93. https://doi.org/10.3846/bme.2018.2142
8. Bolton RJ, Hand DJ (2002) Statistical fraud detection: a review. Stat Sci 17(3):235–255. https://doi.org/10.1214/ss/1042727940
9. Bommarito MJ, Katz DM (2023) GPT takes the bar exam. SSRN preprint. https://doi.org/10.2139/ssrn.4314839
10. Bovens M (2007) Analysing and assessing accountability: a conceptual framework. Eur Law J 13(4):447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x
11. Brynjolfsson E, McAfee A (2014) The second machine age: work, progress, and prosperity in a time of brilliant technologies. WW Norton, New York
12. Busuioc M (2021) Accountable artificial intelligence: holding algorithms to account. Public Adm Rev 81(5):825–836. https://doi.org/10.1111/puar.13293
13. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM-CSUR 41(3):1–58. https://doi.org/10.1145/1541880.1541882
14. Commerford BP, Dennis SA, Joe JR, Ulla JW (2022) Man versus machine: complex estimates and auditor reliance on artificial intelligence. J Accounting Res 60(1):171–201. https://doi.org/10.1111/1475-679X.12407
15. Commerford BP, Eilifsen A, Hatfield RC, Holmstrom KM, Kinserdal F (2024) Control issues: how providing input affects auditors' reliance on artificial intelligence. Contemp Acc Res 41(4):2134–2162. https://doi.org/10.1111/1911-3846.12974
16. Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114–126. https://doi.org/10.1037/xge0000033
17. Dwivedi YK, Hughes L, Ismagilova E et al (2021) Artificial intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int J Inf Manag 57:101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002
18. Eulerich M, Sanatizadeh A, Vakilzadeh H, Wood DA (2024) Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing industries. Rev Acc Stud 29:2318–2349. https://doi.org/10.1007/s11142-024-09833-9

19. Frey CB, Osborne MA (2017) The future of employment: how susceptible are jobs to computerisation? Technol Forecast Soc Chang 114:254–280. https://doi.org/10.1016/j.techfore.2016.08.019
20. Glikson E, Woolley AW (2020) Human trust in artificial intelligence: review of empirical research. Acad Manag Ann 14(2):627–660. https://doi.org/10.5465/annals.2018.0057
21. Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1):121–127. https://doi.org/10.1136/amiajnl-2011-000089
22. Grønsund T, Aanestad M (2020) Augmenting the algorithm: emerging human-in-the-loop work configurations. J Strategic Inform Syst 29(2):101614. https://doi.org/10.1016/j.jsis.2020.101614
23. Gu H, Schreyer M, Moffitt K, Vasarhelyi MA (2024) Artificial intelligence co-piloted auditing. Int J Acc Inform Syst 54:100698. https://doi.org/10.1016/j.accinf.2024.100698
24. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM-CSUR 51(5):93. https://doi.org/10.1145/3236009
25. IAASB (2009) International Standard on Auditing 230: audit documentation. International Auditing and Assurance Standards Board, New York
26. IAASB (2025) International Standard on Auditing 240 (Revised): the auditor's responsibilities relating to fraud in an audit of financial statements. International Auditing and Assurance Standards Board, New York
27. Issa H, Sun T, Vasarhelyi MA (2016) Research ideas for artificial intelligence in auditing: the formalization of audit and workforce supplementation. J Emerg Technol Acc 13(2):1–20. https://doi.org/10.2308/jeta-10511
28. Janvrin D, Bierstaker J, Lowe DJ (2008) An examination of audit information technology use and perceived importance. Acc Horizons 22(1):1–21. https://doi.org/10.2308/acch.2008.22.1.1
29. Jarrahi MH (2018) Artificial intelligence and the future of work: human–AI symbiosis in organizational decision making. Bus Horiz 61(4):577–586. https://doi.org/10.1016/j.bushor.2018.03.007
30. Keding C (2021) Understanding the interplay of artificial intelligence and strategic management: four decades of research in review. Manage Rev Q 71(1):91–134. https://doi.org/10.1007/s11301-020-00181-x
31. Kellogg KC, Valentine MA, Christin A (2020) Algorithms at work: the new contested terrain of control. Acad Manag Ann 14(1):366–410. https://doi.org/10.5465/annals.2018.0174
32. Kokina J, Davenport TH (2017) The emergence of artificial intelligence: how automation is changing auditing. J Emerg Technol Acc 14(1):115–122. https://doi.org/10.2308/jeta-51730
33. Lakis V, Masiulevičius A (2017) Limitations and application possibilities of the monetary unit audit sampling method. Ekonomika 96(2):131–145. https://doi.org/10.15388/Ekon.2017.1.10668

34. Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. MIS Q 45(3):1501–1526. https://doi.org/10.25300/MISQ/2021/16564
35. Lipton ZC (2018) The mythos of model interpretability. Queue 16(3):31–57. https://doi.org/10.1145/3236386.3241340
36. Lo Piano S (2020) Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanit Social Sci Commun 7:9. https://doi.org/10.1057/s41599-020-0501-9
37. Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: people prefer algorithmic to human judgment. Organ Behav Hum Decis Process 151:90–103. https://doi.org/10.1016/j.obhdp.2018.12.005
38. Loughran M (2010) Auditing for dummies. Wiley, Hoboken, NJ
39. National Audit Office (2001) A practical guide to sampling. National Audit Office, London
40. Ngai EWT, Hu Y, Wong YH, Chen Y, Sun X (2011) The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. Decis Support Syst 50(3):559–569. https://doi.org/10.1016/j.dss.2010.08.006
41. O'Leary DE, O'Keefe RM (1997) The impact of artificial intelligence in accounting work: expert systems use in auditing and tax. AI Soc 11(1–2):36–47. https://doi.org/10.1007/BF02812437
42. OpenAI (2023) GPT-4 technical report. arXiv preprint. https://arxiv.org/abs/2303.08774
43. Pang G, Shen C, Cao L, van den Hengel A (2021) Deep learning for anomaly detection: a review. ACM-CSUR 54(2):38. https://doi.org/10.1145/3439950
44. Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. Hum Factors 39(2):230–253. https://doi.org/10.1518/001872097778543886
45. Power M (1997) The audit society: rituals of verification. Oxford University Press, Oxford
46. Raisch S, Krakowski S (2021) Artificial intelligence and management: the automation–augmentation paradox. Acad Manage Rev 46(1):192–210. https://doi.org/10.5465/amr.2018.0072
47. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x
48. Scarbrough H, Chen Y, Patriotta G (2024) The AI of the beholder: intra-professional sensemaking of an epistemic technology. J Manage Stud 62(5):1885–1913. https://doi.org/10.1111/joms.13065
49. Schreyer M, Baumgartner M, Ruud TF, Borth D (2022) Artificial intelligence in internal audit as a contribution to effective governance — deep-learning enabled detection of anomalies. Expert Focus 2022(1):45–50
50. Sclar M, Choi Y, Tsvetkov Y, Suhr A (2024) Quantifying language models' sensitivity to spurious features in prompt design. In: Proceedings of the International Conference on Learning Representations (ICLR 2024). https://arxiv.org/abs/2310.11324

51. Sebele F (2023) The application of artificial intelligence in external auditing. Int J Res Bus Social Sci 12(9):496–512. https://doi.org/10.20525/ijrbs.v12i9.2737
52. Shrestha YR, Ben-Menahem SM, von Krogh G (2019) Organizational decision-making structures in the age of artificial intelligence. Calif Manag Rev 61(4):66–83. https://doi.org/10.1177/0008125619862257
53. Skitka LJ, Mosier KL, Burdick M (1999) Does automation bias decision-making? Int J Hum Comput Stud 51(5):991–1006. https://doi.org/10.1006/ijhc.1999.0252
54. Skitka LJ, Mosier K, Burdick MD (2000) Accountability and automation bias. Int J Hum Comput Stud 52(4):701–717. https://doi.org/10.1006/ijhc.1999.0349
55. Susskind R, Susskind D (2015) The future of the professions: how technology will transform the work of human experts. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198713395.001.0001
56. Sutton SG, Holt M, Arnold V (2016) The reports of my death are greatly exaggerated — artificial intelligence research in accounting. Int J Acc Inform Syst 22:60–73. https://doi.org/10.1016/j.accinf.2016.07.005
57. Umar H, Mayangsari S, Moffitt K (2021) Reciprocal use of artificial intelligence in audit assignments. J Acc Bus Finance Res 11(1):9–20. https://doi.org/10.20448/2002.111.9.20
58. Vasarhelyi MA, Kogan A, Tuttle BM (2015) Big data in accounting: an overview. Acc Horizons 29(2):381–396. https://doi.org/10.2308/acch-51071
59. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). https://arxiv.org/abs/1706.03762
60. Wang J, Wang J, Athiwaratkun B, Zhang C, Zou J (2025) Mixture-of-agents enhances large language model capabilities. In: Proceedings of the International Conference on Learning Representations (ICLR 2025). https://openreview.net/pdf?id=h0ZfDIrj7T
61. Wassie FA, Lakatos LP (2024) Artificial intelligence and the future of the internal audit function. Humanit Social Sci Commun 11:386. https://doi.org/10.1057/s41599-024-02905-w
62. West J, Bhattacharya M (2016) Intelligent financial fraud detection: a comprehensive review. Computers Secur 57:47–66. https://doi.org/10.1016/j.cose.2015.09.005
63. Wilson HJ, Daugherty PR (2018) Collaborative intelligence: humans and AI are joining forces. Harvard Business Rev 96(4):114–123
64. Yao L (2024) Application of artificial intelligence technology in enterprise financial audit. In: 2024 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), pp 1–5. https://doi.org/10.1109/ICICACS60521.2024.10498966