AuditCopilot: Leveraging LLMs for Fraud Detection in Double-Entry Bookkeeping

Md Abdul Kadir[1,2] **Sai Suresh Macharla Vasu**[1,*] **Siddharth S. Nair**[1,*] **Daniel Sonntag**[1,2]
[1] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[2] Oldenburg University, Oldenburg, Germany
{abdul.kadir, sai_suresh.macharla_vasu, sidharth.nair, daniel.sonntag}@dfki.de

Abstract

Auditors rely on Journal Entry Tests (JETs) to detect anomalies in tax-related ledger records, but rule-based methods generate overwhelming false positives and struggle with subtle irregularities. We investigate whether large language models (LLMs) can serve as anomaly detectors in double-entry bookkeeping. Benchmarking SoTA LLMs such as LLaMA and Gemma on both synthetic and real-world anonymized ledgers, we compare them against JETs and machine learning baselines. Our results show that LLMs consistently outperform traditional rule-based JETs and classical ML baselines, while also providing natural-language explanations that enhance interpretability. These results highlight the potential of **AI-augmented auditing**, where human auditors collaborate with foundation models to strengthen financial integrity.

1 Introduction

Financial auditors play a critical role in detecting unusual transactions and anomalies that could indicate errors, fraud, or tax evasion in a company's books. Ensuring that such irregularities are identified is vital for maintaining trust in financial statements Wang et al. [2024]. However, auditing large volumes of accounting data is extremely challenging; current manual and rule-based processes are often inefficient and error-prone, meaning auditors can miss important red flags. Standard Journal Entry Tests (JETs) Droste and Tritschler [2024], which rely on predefined rules and known patterns, tend to flag a huge number of transactions. These full population tests catch only anticipated anomalies Herreros-Martínez et al. [2025] and flood auditors with alerts, many of which are false positives requiring time-consuming review. In practice, auditors still rely heavily on sampling and human labor Wang et al. [2024], which risks overlooking subtle irregularities in today's era of massive, complex datasets Herreros-Martínez et al. [2025], Wang et al. [2024]. This gap between growing data complexity and the limits of traditional audit techniques underscores the need for more intelligent, automated anomaly detection tools in auditing.

A journal entry, in accounting terms, represents the basic record of a financial transaction, typically involving a debit and a credit entry across one or more accounts. Each entry is enriched with attributes such as transaction date, posting ID, currency, user responsible for entry, tax rates, amounts, and free-text descriptions. These heterogeneous features, ranging from numerical to categorical to textual, make journal entries a natural testbed for anomaly detection systems operating in multi-type data environments. Both the synthetic and anonymized datasets share a common feature space that includes these core attributes, facilitating consistent modeling and evaluation across domains.

In response, researchers have explored machine learning (ML) to enhance anomaly detection in accounting data. Unsupervised algorithms, such as clustering and Isolation Forest, can sift through journal entries to pinpoint outliers that deviate from normal patterns Herreros-Martínez et al. [2025]. Recent work on synthetic general ledger data demonstrated that these methods can improve detection

[*] Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

performance beyond rule-based JETs Gronewald et al. [2024]. Still, there are limitations: traditional ML models typically operate on structured features and may lack the contextual understanding needed to differentiate truly risky anomalies from benign outliers.

Meanwhile, the advent of **Large Language Models (LLMs)** has opened up new possibilities for auditing. These foundation models, exemplified by GPT-4 and similar, possess powerful natural language understanding and reasoning capabilities. Early studies suggest that LLMs could serve as AI auditors or co-pilots, working alongside humans to analyze financial data Gu et al. [2023]. Recent advances also indicate that LLMs can be leveraged for **anomaly detection**, such as SigLLM for time-series Alnegheimish et al. [2024] and AnoLLM for tabular records Tsai et al. [2025], both showing competitive performance without heavy feature engineering.

Of course, deploying LLMs in high-stakes audit applications must be done carefully. By nature, LLMs can sometimes produce inconsistent answers, show biases, or hallucinate. Recognizing this, the community has begun to audit the auditors with approaches such as AuditLLM Amirizaniani et al. [2024a] and LLMAuditor Amirizaniani et al. [2024b]. These highlight that while LLMs are powerful, governance and oversight are necessary when using them as decision-support tools in domains like auditing.

In this paper, we leverage general-purpose LLMs for anomaly detection on tax-related ledger data, combining data-driven detection with natural-language explainability using structured input output schema making a scalable method. We evaluate this on two datasets: a large real-world anonymized general ledger obtained from project partners and a synthetic dataset from Gronewald et al. [2024]. By testing on both, we ensure our method works for practical, messy data and benchmark scenarios. Our results show that LLMs can achieve detection performance comparable to classic ML algorithms, while simultaneously providing rich, interpretable explanations that can assist auditors. This represents an encouraging step toward **AI-augmented auditing**, where human auditors and AI systems collaborate to ensure financial integrity.

2 Method

We use prompt tuning of a large language model to decide whether a posting ID (journal-entry (JE) grouping) is fraudulent or normal and to emit a concise explanation. Model weights are never updated; behaviour is controlled by an instruction prompt and a structured input schema. The method operates under weak/no supervision (analogous to Isolation Forest), while natively ingesting heterogeneous numeric, categorical, and textual JE fields without heavy feature engineering. We find the benefits of this method to be multi-fold, grounding explanations are key to establishing trust in financial AI models for non-technical stakeholders, and prompt tuning is a fairly straightforward modification in simple natural language, leading to behaviour modifications in LLMs' decision making, hence giving more control to end-users in adjusting the model's behaviour for specifc use cases.

Addition of contextual information Contextual information is important in delicate decision-making scenarios such as financial auditing, even for human auditors. Motivated by this fact, we augment the LLM prompt by adding summary statistics and percentile information for numerical features such as transaction amounts, and adding frequency of occurrences for categorical features, such as user ID of accountants. As described earlier, when ingesting fine-grained data per transaction instead of posting ID, we find performance gains in our method when adding global dataset statistics, as well as grounding of the explanations with reduced hallucinations, this is corroborated by AD-LLM Yang et al. [2024] shows prompt augmentation with semantic context improves discriminatory performance. Additionally, we add more signal to the prompt by adding the decision of an Isolation Forest Liu et al. [2008] classifier into the prompt. Empirically, we found this improved performance wrt the number of False Positives (FP) registered by our method, effectively reducing the burden of their reevaluation by human auditors, saving time and increasing efficiency.

Prompt specifics Prompts are broken down into system and transaction-specific parts. System prompts include general guidelines as well as the contextual information discussed above. Each instance (posting ID/transaction) is provided as a compact JSON-like record with multi-type fields (date, posting_id, currency, user, tax rates, amounts, account codes, memo text). The model returns a strict JSON object with `anomaly` $\in \{0, 1\}$ and an `explanation`.

3 Dataset and Experiments

The task is framed as **unsupervised fraud detection in multi-type journal entry data**. Labels for fraudulent activity are scarce and often unavailable in practice, making supervised learning infeasible

at scale. Instead, we aim to detect anomalous entries by modeling irregular patterns across structured and unstructured fields to reflect real-world auditing conditions.

Synthetic JE dataset This is a **simulated accounting dataset** introduced by Gronewald et al. [2024], designed to emulate the day-to-day book-keeping practices of a medium-sized enterprise. Features in each entry include a unique posting ID, date and time of posting and of the transaction, Credit/Debit (CD) flags, tax rate, Account ID etc.

Anonymized JE dataset We benchmark our method on real-world anonymized accounting data from our industry partner. Unlike typical double-entry bookkeeping systems where posting IDs link balanced debit-credit cycles, this dataset lacks such identifiers due to anonymization, breaking the usual transactional structure. We therefore treat it as a transaction-level fraud detection problem. To compensate for missing context, we enrich the LLM prompt with bookkeeping heuristics, guidelines, and summary statistics, which proves essential to outperform traditional ML baselines.

Privacy is an important aspect when dealing with confidential information such as accounting data, hence preference was given to open weight LLMs (LLaMa AI [2025], Gemma Team et al. [2024]) over proprietary LLM API endpoints such as (GPT 4o OpenAI [2024]) when benchmarking our method.

Real world (anonymized) and synthetic JE data has been used for our tests. For the real-world accounting data, due to the non-availablity of labels, pseudo-labeling was done purely for evaluations by employing Journal Entry Testing (JET) practices verified by the client involved in the project. JET is a rule-based method used in practice to filter out outliers for further scrutiny by auditors. Control over the anomaly rate is possible in the case of Synthetic data, where we present detailed results for a randomly sampled subset of 5000 posting ID's with a 1% anomaly rate i.e. 50 anomalous entries. For synthetic data, post pseudo labeling with JET, we get $\sim 6\%$ anomaly rate.

4 Results and Discussion

Our method outperforms the baselines provided on the Synthetic data Gronewald et al. [2024] as well as on a real-world anonymised dataset. We use Isolation forest Liu et al. [2008] as our baseline, due to its theoretical advantages in modeling tabular data Grinsztajn et al. [2022], additionally, it is an unsupervised anomaly detection method, hence aligning with the proposed task in Section 3.

Synthetic dataset. Table 1 summarizes results on the synthetic benchmark from Gronewald et al. [2024]. The traditional JET baseline achieves reasonable recall (0.90) but suffers from very high false positives (FP=942), highlighting its limited precision in practice. Isolation Forest, substantially improves over JET with higher precision (0.61) and near-perfect recall (0.98), reducing false positives by an order of magnitude (FP=169).

Among the LLMs, Mistral-8B Team [2024] delivers the strongest overall performance, reaching the best F1 score (0.94), the highest precision (0.90), and the lowest false positives (FP=12) while maintaining very high recall (0.98). Gemma-7B achieves the highest recall (0.99, FN=0), but at a moderate precision cost (0.71). GPT-5-mini OpenAI [2025] also attains perfect recall (FN=0) but still produces several hundred false positives (FP=466). Other models, such as Llama-3.1-8B and Gemma-2B, show balanced but weaker trade-offs. Overall, these results indicate that modern LLMs, particularly Mistral-8B, can match or surpass traditional ML baselines on synthetic accounting data, achieving both low error rates and interpretable natural-language outputs.

Prompt Ablation results Table 2 shows that the full *AuditCopilot* prompt yields the most balanced performance, with Gemma-2B achieving the best recall (0.98, FN=6) and Gemma-7B the highest precision (0.89, FP=32). Removing *Isolation Forest* drastically increases false positives (e.g., Gemma-7B: 32$\rightarrow$1973), while removing *Statistics* collapses recall across all models (e.g., Gemma-2B: FN 6$\rightarrow$306). These findings indicate that Statistics are key for recall, while Isolation Forests are crucial for precision, and both are necessary for reliable anomaly detection.

5 Conclusion

Our results demonstrate that prompt-engineered LLMs, when combined with Isolation Forest scores, can outperform traditional JETs and ML baselines by offering both strong anomaly detection and natural-language rationales. While challenges remain around false positives, cost, and real-world deployment, this study highlights the promise of AI-augmented auditing where accuracy and interpretability are jointly optimized. Importantly, we provide the first evidence that LLMs can be applied to tax-related ledger data, a high-stakes domain where transparency and trust are critical. Our findings suggest that combining LLMs with classical anomaly signals offers a practical path forward,

Table 1: Anomaly detection results on the Synthetic dataset Gronewald et al. [2024], using the prompt template in Fig. 2. Best value per column in **bold**; lowest for FP/FN.

Dataset	Method Group	Method	Precision ↑	Recall ↑	F1 ↑	TP ↑	FP ↓	FN ↓	TN ↑
Synthetic Dataset	Traditional	JET	0.53	0.90	0.50	**50**	942	**0**	4008
	ML Baseline	Isolation Forest	0.61	0.98	0.68	**50**	169	**0**	4781
	LLMs	Gemma 2B	0.53	0.92	0.53	49	685	1	4265
		Gemma 7B	0.71	**0.99**	0.79	**50**	68	**0**	4882
		Llama 3.1 8B	0.67	0.88	0.73	39	78	11	4872
		Mistral 8B	**0.90**	0.98	**0.94**	48	**12**	2	**4938**
		Mistral Small 22B	0.53	0.92	0.52	49	711	1	4239
		GPT-5-mini	0.55	0.95	0.56	**50**	466	**0**	4484

Table 2: Results on Anonymised Dataset, along with Prompt ablations; Best values per column in **bold**. AuditCopilot is the prompt variant shown in Fig. 1, *w/o IF* is the prompt without Isolation Forest results but including statistics; *w/o Stats,IF* is prompt variant without global Statistics or Isolation Forest results

Dataset	Prompt Variant	Method	Precision ↑	Recall ↑	F1 ↑	TP ↑	FP ↓	FN ↓	TN ↑
Private Dataset	-	Isolation Forest	0.30	0.96	0.46	315	719	14	3952
	AuditCopilot	Gemma-2B	0.30	**0.98**	0.46	**323**	740	**6**	3931
		Mistral-8B	0.39	0.90	0.54	295	468	34	4203
		Gemma-7B	**0.89**	0.78	**0.83**	256	**32**	73	**4639**
		Llama-3.1-8B	0.17	0.86	0.28	282	1369	47	3302
	w/o IF	Gemma-2B	0.32	0.79	0.46	259	543	70	4128
		Gemma-7B	0.14	0.90	0.24	311	1973	34	2982
		Llama-3.1-8B	0.18	0.81	0.29	267	1232	62	3439
	w/o Stats, IF	Gemma-2B	0.07	0.07	0.07	23	290	306	4381
		Gemma-7B	0.41	0.22	0.28	71	104	258	4567
		Llama-3.1-8B	0.07	0.35	0.11	112	1668	217	3003

balancing precision, recall, and explainability. Future work will extend these insights to larger, more diverse datasets and explore robustness against prompt variation and model updates.

Limitations

Data scope and labels. Our evaluation is limited to one anonymized tax-related ledger and a synthetic benchmark Gronewald et al. [2024]. Real-world ledgers differ across industries, ERP systems, and jurisdictions, and ground-truth labels are scarce. Pseudo-labels based on JETs may inherit rule-based biases, while synthetic anomalies are simulator-defined. Thus, external validity remains to be demonstrated with larger, multi-source datasets and independent audit labels.

Model stability and explainability. LLM behavior is prompt-sensitive and model-dependent: small changes in phrasing or sampling can shift results, and we observe disagreement across families (e.g., Gemma vs. Mistral). Although our AuditCopilot design reduces variance, stability under paraphrasing, adversarial inputs, and model updates is not guaranteed. Likewise, natural-language explanations may not always faithfully reflect decision boundaries, even if auditors find them useful.

Deployment and governance. While our method shows strong experimental performance, using LLMs in financial auditing raises safety, privacy, and compliance challenges. Risks include hallucinations, automation bias, prompt injection via free-text fields, and lack of reproducibility as model weights evolve. We stress that our approach should be viewed as *decision support*, requiring human oversight, access controls, and governance frameworks before any real-world deployment.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IW24006, and by the Federal Ministry of Research, Technology and Space under grant 16IS23064; it has also been supported by the Ministry for Science and Culture of Lower Saxony (MWK), the Endowed Chair of Applied Artificial Intelligence, Oldenburg University, and DFKI, with industry advisory support from DATEV eG.

References

M. AI. Llama 3.1. https://huggingface.co/meta-llama/Llama-3.1-8B, 2025. Model card, accessed: 2025-11-16.

S. Alnegheimish, L. Nguyen, L. Berti-Equille, and K. Veeramachaneni. Can large language models be anomaly detectors for time series? In *2024 IEEE International Conferencze on Data Science and Advanced Analytics (IEEE DSAA)*. IEEE, 2024.

M. Amirizaniani, E. Martin, T. Roosta, A. Chadha, and C. Shah. Auditllm: A tool for auditing large language models using multiprobe approach, 2024a. URL https://arxiv.org/abs/2402.09334.

M. Amirizaniani, J. Yao, A. Lavergne, E. S. Okada, A. Chadha, T. Roosta, and C. Shah. Llmauditor: A framework for auditing large language models using human-in-the-loop, 2024b. URL https://arxiv.org/abs/2402.09346.

K. C. Droste and J. Tritschler. *Journal Entry Testing*. IDW Verlag, Düsseldorf, 2 edition, 2024. ISBN 978-3-8021-2934-6.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

J. Gronewald, A. M. Rombach, S. Stephan, and P. Fettke. Anomaly detection in general ledger data: Results from a hybrid approach, 2024.

H. Gu et al. Artificial intelligence co-piloted auditing, 2023. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4444763. SSRN preprint 4444763.

A. Herreros-Martínez, R. Magdalena-Benedicto, J. Vila-Francés, A. J. Serrano-López, S. Pérez-Díaz, and J. J. Martínez-Herráiz. Applied machine learning to anomaly detection in enterprise purchase processes: A hybrid approach using clustering and isolation forest. *Information*, 16(3), 2025. ISSN 2078-2489. doi: 10.3390/info16030177. URL https://www.mdpi.com/2078-2489/16/3/177.

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.

OpenAI. Gpt-4o, 2024. URL https://chat.openai.com. Large language model used for anomaly detection. Version: GPT-4o, Accessed: 2025-05-24.

OpenAI. Gpt-5 mini. https://platform.openai.com/docs/models/gpt-5-mini, 2025. Model variant of GPT-5 family.

G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

M. A. Team. Ministral 8b (instruct-2410). https://huggingface.co/mistralai/Ministral-8B-Instruct-2410, 2024. Model card, access date: 2025-11-16.

C.-P. Tsai, G. Teng, P. Wallis, and W. Ding. AnoLLM: Large language models for tabular anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7VkHffT5X2.

R. Wang, J. Liu, W. Zhao, S. Li, and D. Zhang. Automating financial statement audits with large language models. In *2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle*, 2024. URL https://openreview.net/forum?id=4MaWVsVb2g.

T. Yang, Y. Nian, S. Li, R. Xu, Y. Li, J. Li, Z. Xiao, X. Hu, R. Rossi, K. Ding, et al. Ad-llm: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*, 2024.

A Related Work

Machine Learning for Anomaly Detection. A large body of research has explored machine learning (ML) to enhance anomaly detection in accounting data. Unsupervised algorithms, such as clustering and Isolation Forest, can sift through journal entries to pinpoint outliers that deviate from normal patterns Herreros-Martínez et al. [2025]. On synthetic general ledger data, these methods improved detection beyond rule-based JETs; notably, Isolation Forest was found to be effective at uncovering fraudulent entries Gronewald et al. [2024]. To aid practitioners, explainability techniques like SHAP have been applied so that flagged anomalies are accompanied by insights about why they were deemed suspicious Herreros-Martínez et al. [2025]. Still, such approaches are limited: traditional ML models operate mainly on structured features, produce risk scores or clusters, and often cannot provide natural-language justifications without additional tooling or domain expertise.

LLMs as Auditing Assistants. The advent of Large Language Models (LLMs) has opened new avenues for auditing. These foundation models possess strong natural language reasoning abilities, making them promising co-pilots for auditors. For example, Gu et al. [2023] fine-tuned a GPT-4 model with chain-of-thought prompting to support several audit tasks, including journal entry testing, and demonstrated how such a system can systematically guide audit procedures. Their work highlights the potential of LLMs to augment efficiency and provide richer insights in auditing contexts.

LLMs for Anomaly Detection. Recent advances indicate that pretrained LLMs can also serve as anomaly detectors. SigLLM Alnegheimish et al. [2024] reformulates time-series sensor data into textual sequences and prompts an LLM to identify irregular patterns, achieving performance comparable to specialized algorithms without task-specific training. Similarly, AnoLLM Tsai et al. [2025] serializes tabular records into text and leverages an LLM to assign anomaly scores, outperforming many conventional models across benchmark datasets. These approaches demonstrate the versatility of LLMs in ingesting heterogeneous data formats and applying contextual reasoning.

Auditing the Auditors. Deploying LLMs in high-stakes audit settings raises concerns about consistency, bias, and hallucinations. To address this, the community has begun developing frameworks to evaluate and govern LLM-based auditors. AuditLLM Amirizaniani et al. [2024a] probes model consistency across query rephrasings, while LLMAuditor Amirizaniani et al. [2024b] automates large-scale auditing with a combination of LLM checks and human-in-the-loop oversight. These works emphasize the importance of governance mechanisms alongside technical advances.

B Prompt Templates

We used two standardized prompt formats: the *Vanilla Prompt* for the private dataset (Figure 1) and the *Synthetic Prompt* for the synthetic dataset (Figure 2).

You are an expert financial auditor specializing in journal entry testing (JET) and fraud detection.

DATASET CONTEXT:

- Total transactions in dataset: {total_transactions}
- Isolation Forest detected {total_if_anomalies} anomalies ({if_anomaly_rate})
- Amount statistics:
  - Mean: {amount_mean}
  - Median: {amount_median}
  - 95th percentile: {amount_q95}
  - 99th percentile: {amount_q99}
  - Range: {amount_min} to {amount_max}
- Payment period max: {payment_period_max} days
- Total users: {total_users}
- Total accounts: {total_accounts}

ANOMALY DETECTION CRITERIA:

1. Unusual amounts: extremely high/low compared to typical transactions
2. Round number bias: large round numbers (1000, 5000, 10000, ...)
3. User behavior: low-volume users handling high-value transactions
4. Timing anomalies: very long or very short payment periods
5. Account patterns: unusual account relationships or concentrations
6. Statistical outliers: values far from normal distribution
7. Business logic violations: transactions that don't make business sense

RESPONSE FORMAT: Respond ONLY with valid JSON in this exact format:

```
{"anomaly": 0, "explanation": "Normal transaction, with reasoning"}
OR
{"anomaly": 1, "explanation": "Specific reason why this transaction is anomalous"}
```

Be conservative – only flag clear anomalies. Provide specific, actionable explanations.

TRANSACTION DATA: {transaction_data}
Isolation Forest Hint: {if_status} (score: {if_score})

ADDITIONAL CONTEXT:

- This user ({user_id}) has {user_tx_count} total transactions
- This amount ({abs_amount}) is at the {amount_percentile}th percentile

Evaluate for anomalies and respond with JSON only.

Figure 1: AuditCopilot prompt template with dataset statistics and Isolation Forest hints

You are an expert financial auditor reviewing a journal entry, which may contain multiple transactions.

For each transaction, the following engineered features are provided:

- `promptly`: 1 = 0–9 days (not triggered), 2 = 10–29 days (triggered), 3 = 30 days (triggered)
- `weekend`: 0 = Mon–Fri (not triggered), 1 = Saturday (triggered), 2 = Sunday (triggered)
- `nwh`: 0 = during working hours (not triggered), 1 = outside working hours (triggered)
- `top_n`: 0 = not in top_n (not triggered), 1 = in top_n (triggered)
- `high_cash`: 0 = not high cash (not triggered), 1 = high cash (triggered)

Flag rules: A flag is considered "triggered" if:

- `promptly` is 2 or 3
- `weekend` is 1 or 2
- `nwh`, `top_n`, or `high_cash` equals 1

Decision rule: Mark `anomaly = 1` only if **two or more flags are triggered** in the entry.

Response format: Respond *only* with a strict JSON object in this format:

```
{
  "anomaly": 0 or 1,
  "confidence": <decimal in [0,1]>,
  "explanation": "<brief text>"
}
```

Terminate your response with `<|endofanalysis|>`.

Figure 2: Synthetic dataset prompt template with engineered flags and rule-based decision criteria.