Accounts Payable automation mechanism based on large language models

Oleksandr Rotar[1]

Опубліковано	Секція	УДК
30.11.2024	Економіка	657.41:004.8

DOI: https://doi.org/10.5281/zenodo.20917048

Abstract. The study is aimed at developing and empirically testing an intelligent invoice processing procedure that integrates field extraction, discrepancy detection, explanation generation, and document routing within a controlled accounting workflow. It is to be seen whether the LLM-based approach gives better preliminary invoice verification than baseline OCR/RPA logic.

It was developed through the results of empirical testing using a simulated corpus of 150 invoices and related control documents. Out of these, 90 had no discrepancies, while 60 were deliberately made to have discrepancies: errors in the amount, missing or wrong purchase order, supplier mismatch, date error, or currency error. Every document was processed by both the LLM-based mechanism and the baseline OCR/RPA approach, after which their outputs were compared with ground-truth annotation. The parameters for evaluation were accuracy, precision, recall, and F1-score; additionally, the LLM explanations were analyzed for factual consistency, clarity, and unsupported claims.

The LLM-based mechanism accurately classified 136 out of 150 documents as opposed to 116 documents correctly classified by the baseline. This translated to an accuracy of 90.7% as opposed to 77.3%, precision of 87.1% as opposed to 73.2%, recall of 90.0% as opposed to 68.3%, and F1-score of 88.5% as opposed to 70.7%. The LLM-based mechanism missed 6 real discrepancies; the baseline missed 19. The overall accuracy of field extraction and determination was 93.8% for the LLM and 87.9% for the baseline. At the same time, it confirmed 8 false risk flags and 8 incorrect explanations, thus underlining the need for human-in-the-loop control.

The scientific novelty lies in the empirical justification of the LLM as a controlled semantic layer. This layer exists between the invoice, ground-truth data, purchase order, approval routing, and manual review. The practical value of the results is the possibility of using the procedure proposed for preliminary screening of risky invoices, reducing missed discrepancies, creating an audit trail, and supporting the work of financial controllers.

Keywords: invoices, financial control, OCR, RPA, purchase order, field extraction, semantic validation, ground-truth annotation, human-in-the-loop, audit trail.

Механізм автоматизації Accounts Payable на основі великих мовних моделей

Анотація. Метою дослідження є розробка та практичне тестування інтелектуальної процедури обробки рахунків-фактур, яка поєднує вилучення деталей, виявлення невідповідностей, генерацію пояснень та маршрутизацію документів у рамках контрольованого бухгалтерського процесу. Це дослідження перевірить, чи забезпечує підхід LLM кращу попередню перевірку рахунків-фактур порівняно з базовою

[1] Senior Software Engineer, MagMutual Insurance Agency, LLC, Atlanta, GA, USA, ORCID: https://orcid.org/0009-0004-7358-899X

логікою OCR/RPA. Методологія базувалася на емпіричному тестуванні змодельованого корпусу зі 150 рахунків-фактур та пов'язаних з ними контрольних документів: 90 документів без розбіжностей та 60 із змодельованими винятками (останні мали різні типи помилок: неправильна сума, відсутнє або неправильне замовлення на купівлю, невідповідність постачальника, помилка дати або валюти). Кожен документ оброблявся механізмом LLM та базовим підходом OCR/RPA; результати порівнювалися з довідковою розміткою. Оцінювання проводилося на основі показників точності, прецизійності, повноти та F1-оцінки; пояснення LLM також перевірялися на відповідність реквізитам, ясність та відсутність непідтверджених тверджень.

Було виявлено, що механізм LLM правильно класифікував 136 зі 150 документів, тоді як базовий підхід – 116. Точність становила 90,7% проти 77,3%, прецизійність – 87,1% проти 73,2%, повнота – 90,0% проти 68,3%, F1-оцінка – 88,5% проти 70,7%. Механізм LLM пропустив 6 реальних розбіжностей, тоді як базовий підхід пропустив 19. Загальна точність вилучення та визначення реквізитів становила 93,8% для LLM та 87,9% для базового рівня. Водночас було зафіксовано 8 хибних позначень ризику та 8 неправильних пояснень, що підтверджує необхідність контролю «людини в циклі».

У цій статті представлено емпірично обґрунтовану LLM як контрольований семантичний шар між рахунком-фактурою, довідковими даними, замовленням на купівлю, маршрутизацією затвердження та ручною перевіркою. Отримані результати можуть бути практично корисними при можливості застосування запропонованої процедури для попереднього відбору ризикованих рахунків-фактур, мінімізації кількості пропущених розбіжностей, формування аудиторського сліду та підтримки роботи фінансового контролера.

Ключові слова: інвойси, фінансовий контроль, OCR, RPA, purchase order, вилучення реквізитів, семантична перевірка, еталонна розмітка, human-in-the-loop, аудиторський слід.

Introduction

Relevance of the problem. The scope of Accounts Payable automation today is the processing of the invoice from the time it is received in the finance department to the time it is paid. This includes the receipt of the invoice from the supplier through different channels, verification of the supplier, matching of the invoice with the purchase order, detection of duplicate payments, and establishment of a proper control trail. The traditional OCR and RPA solutions will reduce some manual work and do much of this but typically rely on relatively stable document structures, fixed templates, and rules that must be predefined.

Large language models change this because they can perform not only data extraction but also content interpretation of a financial document, explanation of discrepancies, and review support for non-standard transactions. Therefore, LLM-based Accounts Payable automation is relevant as an organisational and accounting mechanism that consolidates invoice processing, internal control, the ERP environment, payment approval, and audit verification.

Analysis of recent research and publications. The study of the automation of financial and accounting processes has developed several related approaches. The first approach treats RPA as a tool that performs repetitive accounting tasks. L. Cooper et al. stress its capability in reducing labour intensiveness in routine operations [1, p. 15]; RPA is associated with the emergence of “digital labour” in accounting by J. Kokina and S. Blanchette [2]; and problems of integration, data quality, and control over the outputs of smart RPA are drawn by M. Gotthardt et al. [3, p. 90]. These studies mainly describe automation as the execution of predefined rules and do not address fully semantic verification of financial documents.

The second approach is smart document interpretation. R. Palm et al. introduced CloudScan as one of the first systems for neuro-network-based invoice analysis [4, p. 406]; P. Lai et al. built a special pipeline for commercial invoice data extraction and processing [5]. The

further evolution of document AI includes BERT, BERTgrid, DocFormer, LayoutLMv3, and models for parsing spatial dependencies, which allow consideration of text, context, visual structure, and spatial relations in semi-structured documents [6, p. 4171; 7; 8, p. 993; 9, p. 4083; 10, p. 330]. This is very important for Accounts Payable because an invoice is not a classic document consisting of a text only. It has requisites; it has the logic of a business transaction and relates to a purchase order plus an approval path afterward.

The third approach relates to large and generative language models. T. Brown et al. showed that LLMs could accomplish tasks in a few-shot learning setting [11]; Y. Fan et al. later confirmed their promise for information extraction and warned against overly optimistic evaluations of generative model outputs [12, p. 6409]. In other words, in the context of Accounts Payable, we should think of an LLM not as an independent substitute for either an accountant or an ERP system but rather as a semantic layer for explanation, classification, exception checking, and human-in-the-loop control that it supports.

It is necessary to consider the organisational dimension of digitalisation. O. Edo et al. demonstrate that the adoption of digital technologies is based not only on technical efficiency but also on perceived usefulness, ease of use, and organisational support [13]. Ukrainian studies also report changes in the professional requirements of accountants. M. Kulnych et al. note the necessity of digital literacy [14, p. 216], while M. Petchenko et al. study practical trends in the digitalisation of accounting in Ukraine [15, p. 105]. However, there is still no integrated model of LLM-based Accounts Payable automation in the existing literature that would unite data extraction, semantic verification, risk control, approval routing, and the creation of an audit trail.

Unresolved issues of the problem. The substantial progress of RPA, document AI, and generative language models aside, the integrated design of an Accounts Payable automation solution has not been resolved. The large language models should be part of the integrated mechanism as a controlled semantic layer between the invoice and the ERP system, as well as three-way matching procedures, approval routing, internal control, and audit. In other words, the large language models should not be implemented as just another isolated text recognition tool.

Aim of the article is to develop an LLM-based mechanism for Accounts Payable automation that combines intelligent extraction of data from financial documents, semantic verification of transactions, risk control, approval routing, and the creation of an audit trail within a controlled accounting process.

Results

The testing of the empirical Accounts Payable automation mechanism was performed on a simulated corpus of 150 invoices: 90 of them were for standard AP cases without discrepancies, and 60 were pre-modelled exceptions. Each of the documents went through the processing by two approaches: the LLM-based mechanism and the baseline OCR/RPA logic. The output of both approaches was compared with the reference markup, which contained correct requisites, compliance status, type of exception, and recommended action.

The first evaluation stage dealt with the general classification of invoices into two processing routes: either approval without exceptions or transfer for manual review. The results of the classification are presented in Table 1.

Table 1

General invoice classification results of the LLM-based mechanism and baseline approach

Indicator	LLM-based mechanism	Baseline OCR/RPA
True positive	54	41

False positive	8	15
False negative	6	19
True negative	82	75
Correctly classified documents	136	116
Incorrectly classified documents	14	34
Share of documents transferred for manual review	41.3%	37.3%

Source: calculated by the author based on the empirical testing of 150 invoices.

The data in Table 1 shows that the LLM-based mechanism correctly classified 136 out of 150 documents, while the baseline approach correctly classified 116 documents. The number of real discrepancies missed was 6 cases for the LLM-based mechanism and 19 cases for the baseline OCR/RPA logic. This tells us that the baseline approach more often approved documents that should have been moved to manual review.

Based on the TP/FP/FN/TN matrix, accuracy, precision, recall, and F1-score were calculated. The summarised quantitative results are presented in Table 2.

Table 2

Quantitative metrics for classification performance

Metric	LLM-based mechanism	Baseline OCR/RPA	Difference
Accuracy	90.7%	77.3%	+13.4 p.p.
Precision	87.1%	73.2%	+13.9 p.p.
Recall	90.0%	68.3%	+21.7 p.p.
F1-score	88.5%	70.7%	+17.8 p.p.

Source: calculated by the author based on the TP/FP/FN/TN matrix of the empirical testing.

The major difference between the two approaches was in recall. The LLM-based mechanism detected 90.0% of all simulated exceptions, whereas the baseline approach detected 68.3%. This holds true for Accounts Payable because a missed discrepancy would create higher operational risk than the misfiling of a document for further review.

The second block of results concerned the accuracy of extracting and determining invoice requisites. For each invoice, nine fields were checked: supplier name, invoice number, document date, amount payable, currency, tax requisites, purchase order number, compliance status, and recommended action. In total, 1,350 values were evaluated. The results are presented in Table 3.

Table 3

Accuracy of extracting and determining requisites by field type

Document field	LLM-based mechanism, correct	LLM-based mechanism, %	Baseline OCR/RPA, correct	Baseline OCR/RPA, %
Supplier name	145	96.7	140	93.3
Invoice number	147	98.0	143	95.3
Document date	141	94.0	139	92.7
Amount payable	144	96.0	141	94.0
Currency	148	98.7	147	98.0
Tax requisites	132	88.0	121	80.7
Purchase order number	139	92.7	126	84.0

Compliance status	136	90.7	116	77.3
Recommended action	134	89.3	113	75.3
Total	1266	93.8	1186	87.9

Source: calculated by the author based on the comparison of output data with the reference markup.

As shown in Table 3, the LLM-based mechanism correctly extracted or determined 1,266 out of 1,350 values, while the baseline approach correctly processed 1,186 values. The highest LLM scores were obtained for currency, invoice number, supplier name, and amount payable. The lowest values were recorded for tax requisites, compliance status, and recommended action. In the baseline approach, the most problematic fields were recommended action, compliance status, and purchase order number.

The third stage of evaluation examined the ability of both approaches to detect specific types of exceptions. The corpus contained 60 documents with discrepancies. Their distribution and detection results are presented in Table 4.

Table 4

Detection of exceptions by discrepancy type

Type of exception	Number in the corpus	Detected by LLM	Detected by baseline
Incorrect amount	14	13	11
Missing or erroneous purchase order	16	14	10
Supplier mismatch	12	10	8
Date error	10	9	7
Currency error	8	8	5
Total	60	54	41

Source: prepared by the author based on the processing of 60 invoices with simulated exceptions.

The LLM-based mechanism correctly detected 54 out of 60 exceptions. Six cases remained undetected: two concerned supplier mismatches, two related to purchase order issues, one concerned an amount error, and one concerned a date error. The baseline approach missed 19 exceptions, most often in cases where the purchase order number was placed in an atypical location or where the discrepancy required several requisites to be compared at the same time.

The fourth block of results concerned the structure of errors. Since one document could contain several types of deviation, the number of recorded errors does not coincide with the number of incorrectly classified documents. The summary is provided in Table 5.

Table 5

Types of errors during invoice processing

Type of error	LLM-based mechanism	Baseline OCR/RPA
Incorrect extraction of a requisite	5	11
Missed discrepancy	6	19
Erroneous risk identification	8	15
Incorrect recommended action	16	37
Incorrect or incomplete explanation	18	Not generated
Overconfident response	7	Not assessed

Source: systematised by the author based on qualitative error analysis.

The major problem for the LLM-based mechanism turned out to be not the extraction of requisites but the quality of the recommended action or explanation. In 16 cases, the recommended action did not fully correspond to the reference decision. In 18 cases, the explanation was incomplete or contained an inaccuracy. The baseline approach suffered from an incorrect recommended action most typically since rule-based logic did not take into account the substantive context of the document.

The fifth block of results concerned the explanations generated by the LLM-based mechanism. The evaluation was conducted according to three criteria: consistency with factual requisites, clarity for a financial controller, and absence of unsupported claims. The results are presented in Table 6.

Table 6

Quality of explanations generated by the LLM-based mechanism

Explanation category	Number	Share
Fully correct explanation	118	78.7%
Partially correct explanation	24	16.0%
Incorrect explanation	8	5.3%
Total	150	100%

Source: assessed by the author according to the criteria of consistency with requisites, clarity, and absence of unsupported claims.

Among the partially correct explanations, 11 contained an incomplete justification of the recommended action, 7 did not take into account one relevant requisite, and 6 described the risk too generally. Among the incorrect explanations, 5 referred to a requisite that was not present in the invoice, while 3 incorrectly explained the reason for transferring the document for manual review.

The sixth stage of evaluation concerned document routing. For Accounts Payable, it is important not only to extract requisites correctly, but also to determine whether a document may be preliminarily approved or should be transferred to the responsible person. The routing results are presented in Table 7.

Table 7

Distribution of documents by routing outcome

Routing outcome	LLM-based mechanism	Baseline OCR/RPA
Correctly transferred for manual review	54	41
Incorrectly transferred for manual review	8	15
Correctly approved without exceptions	82	75
Incorrectly approved despite an exception	6	19
Total	150	150

Source: calculated by the author based on the comparison of processing routes with the reference markup.

The LLM-based mechanism transferred 62 documents for manual review, of which 54 actually contained exceptions. The baseline approach transferred 56 documents for manual review, but only 41 of them contained real discrepancies. At the same time, the baseline approach incorrectly approved 19 documents with exceptions, which represents the riskiest type of error in the AP process.

Based on the empirical testing, a functional model of the LLM-based mechanism within the Accounts Payable process was developed. It does not involve automatic payment execution,

but ends with the formation of a document status, an explanation, and a route for further review. The resulting model is presented in Table 8.

Table 8

Functional model of the LLM-based Accounts Payable automation mechanism based on the testing results

Stage of the AP process	Input data	Output of the LLM-based mechanism	Control action
Invoice receipt	Invoice text	Preliminary reading of the document	Verification of file completeness
Requisite extraction	Invoice fields	Supplier, date, amount, currency, purchase order	Comparison with reference data
Semantic verification	Invoice and control data	Determination of compliance or exception	Review of risky documents
Recommendation generation	Verification result	"Approve" or "manual review"	Decision by the responsible person
Audit trail	Requisites, status, explanation	Recording of the verification result	Preservation for further control

Source: developed by the author based on the empirical testing of the LLM-based mechanism and process modelling of Accounts Payable.

Results show that the LLM-based mechanism achieved higher accuracy, precision, recall, and F1-score than the baseline OCR/RPA logic. The most visible difference was in the ability not to miss real exceptions: while the LLM-based mechanism failed to detect 6 discrepancies, the baseline approach missed 19. Testing also revealed residual risks: erroneous risk labels, partially incorrect explanations, and overconfident responses. So, the result of the study is not a model of autonomous payment approval but a mechanism for preliminary invoice review, explanation, and routing within a human-in-the-loop Accounts Payable process.

Conclusions

1. It was established that the Accounts Payable automation mechanism based on a large language model provided higher invoice classification accuracy than the baseline OCR/RPA logic. The LLM-based mechanism correctly classified 136 out of 150 documents, whereas the baseline approach correctly classified 116 documents. Overall accuracy was 90.7% compared with 77.3%.
2. It was demonstrated that the main advantage of the LLM-based mechanism lies in discrepancy detection. Recall reached 90.0% for the LLM-based mechanism and 68.3% for the baseline OCR/RPA approach. This means that the LLM-based mechanism missed 6 real exceptions, while the baseline approach missed 19, which is significant for reducing operational risk in Accounts Payable.
3. It was found that the LLM-based mechanism performs more effectively with requisites that require contextual verification, particularly compliance status, purchase order, and recommended action. The overall accuracy of requisite extraction and determination was 93.8% for the LLM-based mechanism and 87.9% for the baseline approach.
4. It was confirmed that an LLM cannot be used as a fully autonomous tool for payment approval. The testing recorded 8 erroneous risk labels, 6 missed discrepancies, and 8 incorrect explanations. Therefore, the appropriate model is one in which the LLM performs preliminary review, generates explanations, and routes documents, while the final decision on risky payments remains with the responsible person.

5. The developed Accounts Payable automation mechanism should be used as a controlled human-in-the-loop process that combines requisite extraction, semantic verification, exception detection, recommendation generation, and preservation of an audit trail. This model is suitable for preliminary automated invoice review, but it requires access control, action logging, and protection of confidential financial data.

Future research should focus on testing the proposed mechanism on real anonymised corporate invoices, assessing its performance in different ERP environments, comparing several LLM architectures, and defining risk thresholds for automated document routing in Accounts Payable practice.

References

1. Cooper, L. A., Holderness, D. K., Jr., Sorensen, T. L., & Wood, D. A. (2019). Robotic process automation in public accounting. *Accounting Horizons, 33*(4), 15-35. https://doi.org/10.2308/acch-52466

2. Kokina, J., & Blanchette, S. (2019). Early evidence of digital labor in accounting: Innovation with robotic process automation. *International Journal of Accounting Information Systems, 35*, Article 100431. https://doi.org/10.1016/j.accinf.2019.100431

3. Gotthardt, M., Koivulaakso, D., Paksoy, O., Saramo, C., Martikainen, M., & Lehner, O. (2020). Current state and challenges in the implementation of smart robotic process automation in accounting and auditing. *ACRN Journal of Finance and Risk Perspectives, 9*(1), 90-102. https://doi.org/10.35944/jofrp.2020.9.1.007

4. Palm, R. B., Winther, O., & Laws, F. (2017). CloudScan: A configuration-free invoice analysis system using recurrent neural networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition* (pp. 406-413). IEEE. https://doi.org/10.1109/ICDAR.2017.74

5. Lai, P., Mohan, A., Kim, S., Chu, J. S. V., Lee, S., Kafle, P., & Wang, P. (2023). Customized information extraction and processing pipeline for commercial invoices. *International Journal of Pattern Recognition and Artificial Intelligence, 37*(09), Article 2354013. https://doi.org/10.1142/S0218001423540137

6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

7. Denk, T. I., & Reisswig, C. (2019). BERTgrid: Contextualized embedding for 2D document representation and understanding. *arXiv*. https://doi.org/10.48550/arXiv.1909.04948

8. Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). DocFormer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 993-1003). https://openaccess.thecvf.com/content/ICCV2021/html/Appalaraju_DocFormer_End-to-End_Transformer_for_Document_Understanding_ICCV_2021_paper.html

9. Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. In *MM '22: Proceedings of the 30th ACM International Conference on Multimedia* (pp. 4083-4091). Association for Computing Machinery. https://doi.org/10.1145/3503161.3548112

10. Hwang, W., Yim, J., Park, S., Yang, S., & Seo, M. (2021). Spatial dependency parsing for semi-structured document information extraction. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 330-343). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.28

11. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

12. Fan, Y., Liu, Y., Yao, Z., Yu, J., Hou, L., & Li, J. (2024). Evaluating generative language models in information extraction as subjective question correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 6409-6417). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.567/

13. Edo, O. C., Ang, D., Etu, E.-E., Tenebe, I., Edo, S., & Diekola, O. A. (2023). Why do healthcare workers adopt digital health technologies? A cross-sectional study integrating the TAM and UTAUT model in a developing economy. *International Journal of Information Management Data Insights, 3*(2), Article 100186. https://doi.org/10.1016/j.jjimei.2023.100186

14. Kulnych, M., Shvorak, A., & Zhylenko, L. (2020). Vprovadzhennia tsyfrovoi hramotnosti v umovakh maibutnikh zmin profesii bukhhaltera [Implementation of digital literacy in the conditions of future changes in the accounting profession]. *Ekonomichnyi Chasopys Skhidnoievropeiskoho Natsionalnoho Universytetu Imeni Lesi Ukrainky, 1*(21), 216-224. https://doi.org/10.29038/2411-4014-2020-01-216-224

15. Petchenko, M., Fomina, T., Balaziuk, O., Smirnova, N., & Luhova, O. (2023). Analiz tendentsii uprovadzhennia tsyfrovizatsii ta dydzhytalizatsii v bukhhalterskyi oblik: Ukrainskyi keis [Analysis of trends in the implementation of digitalization and digitization in accounting: The Ukrainian case]. *Financial and Credit Activity: Problems of Theory and Practice, 1*(48), 105-113. https://doi.org/10.55643/fcaptp.1.48.2023.3951