Automated Financial Data Extraction Using Large Language Models: An Application of OpenAI Apis

Mohd Muhaimin Chuweni[1], Sharifalillah Nordin[2], Jasrul Nizam Ghazali[3*], Mohamad Norzamani Sahroni[3], Mohd Azry Abdul Malik[4]

[1]Finance Division, International Islamic University Malaysia (IIUM), Kuala Lumpur, Malaysia

[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia

[3]Pusat Asasi, Universiti Teknologi MARA (UiTM) Cawangan Selangor, Kampus Dengkil, Malaysia

[4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Machang, Kelantan, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.10200120

Received: 12 February 2026; Accepted: 18 February 2026; Published: 26 February 2026

ABSTRACT

Financial data extraction, traditionally a manual and labour-intensive process, is being revolutionized by artificial intelligence (AI) and machine learning (ML). However, understanding financial documents remains a significant challenge for individuals without specialized financial knowledge due to complex terminology and concepts. This study addresses this gap by designing, developing, and evaluating an AI-powered financial data extraction system tailored for non-financial individuals. The system integrates Optical Character Recognition (OCR) for text extraction from document images statements, invoices, receipts) and leverages the OpenAI platform's advanced Natural Language Processing (NLP) capabilities to organize, interpret, and explain financial information in a user-friendly manner. A Waterfall development methodology was employed, encompassing requirements gathering via questionnaires with target users, system architecture design, implementation using Python libraries and OpenAI API, and rigorous testing, including functionality tests and user evaluations. Results from functionality testing confirmed the system's ability to accurately process various document types. User evaluation, involving finance staff assessing the system's potential for non-expert users, yielded overwhelmingly positive feedback, with high ratings for accuracy, usability, efficiency, and the significant impact of AI/ML integration in enhancing the depth and speed of analysis. The findings demonstrate the system's potential to improve financial literacy and empower individuals in managing personal finances by making complex financial data more accessible and understandable.

Keywords: Financial Data Extraction, OpenAI, Artificial Intelligence, Natural Language Processing, Optical Character Recognition

INTRODUCTION

The management and interpretation of financial data are fundamental to economic decision-making, yet the process has historically been manual, time-consuming, and prone to error. The advent of artificial intelligence (AI) and machine learning (ML) has introduced transformative potential, automating many tasks associated with extracting relevant economic information from diverse sources like financial statements, invoices, and receipts. Technologies such as Optical Character Recognition (OCR) for digitizing text and Natural Language Processing (NLP) for understanding unstructured text work synergistically to enhance efficiency and accuracy in financial data handling.

Despite these technological advancements, a significant barrier persists for non-financial individuals attempting to comprehend financial documents. Financial statements, the lifeblood of organizational assessment, contain crucial insights but often employ specialized terminology, complex concepts, and adhere to regulatory standards that can be bewildering for those without prior financial knowledge or experience. This complexity is exacerbated by reliance on estimates, subjective judgments, and the limitations of traditional metrics (Sherman, 2016). Interpreting ratios and making informed decisions becomes daunting without a solid grasp of accounting principles (Guay, 2016). Furthermore, factors like inadequate training for those preparing reports (Tshehla, 2022) and the inherent complexity hindering investor analysis (Chakraborty, 2019)

underscore the pervasive nature of this challenge. While resources like introductory courses or simplified reports exist (Veal, 2005), there remains a gap for accessible, technology-driven tools specifically designed to help laypersons navigate and understand their *personal* financial documents.

The significance of this project lies in its potential to revolutionize personal financial management by making complex information accessible and understandable. By leveraging advanced AI, the system aims to streamline data handling, reduce errors, enhance financial literacy, and empower users to make more informed decisions regarding budgeting, expense tracking, and planning. The scope involves developing an AI-driven, user-friendly platform integrating OCR and OpenAI's NLP for extracting and interpreting data from user-uploaded financial document images, with evaluation focused on accuracy, efficiency, user-friendliness, and adherence to security best practices. The expected outcome is a robust and transformative automated system that promotes financial literacy and informed decision-making among its target users.

LITERATURE REVIEW

The field of automated data extraction has evolved significantly, moving from manual processes to sophisticated AI-driven techniques. OCR technology is foundational, enabling the conversion of images or printed documents into machine-readable text, a critical first step in digitizing financial records (J. Memon et al., 2020). NLP techniques are then employed to analyse and understand this text, extracting meaningful information and context (Khurana et al., 2023). The integration of OCR and NLP offers a robust approach to handle the complexities of text extraction from diverse document sources, often involving post-processing pipelines to refine accuracy (Rakshit et al., 2023). Text Mining and Information Retrieval further enable the discovery of patterns and relevant information within large volumes of unstructured text, proving valuable in analysing natural language documents (Salloum et al., 2018).

Within the finance and accounting domain, data extraction is pivotal for reporting, analysis, and strategic decision-making. The exponential growth of financial data has spurred the application of text-mining technologies for tasks like financial forecasting, banking analysis, and corporate finance research (Gupta et al., 2020). Specific techniques have also been developed for automating tasks like bank statement reconciliation, ranging from standardized parsers to statistical and data mining approaches, highlighting the drive towards accuracy and efficiency in financial record management (Xing, 2021). Recent advancements in AI, particularly from research labs like OpenAI, have significantly pushed the boundaries of NLP. The Generative Pre-Trained Transformer (GPT) series, built on the transformer architecture, demonstrates remarkable capabilities in understanding and generating human-like text, impacting numerous NLP tasks (Yenduri et al., 2023). These large language models (LLMs) offer powerful tools for interpreting complex information, including financial language (Gruetzemacher, 2022). Related ML paradigms like supervised learning, which relies on labelled data (Cunningham et al., 2008), and reinforcement learning (RL), where agents learn through interaction and feedback (Fahad Mon et al., 2023), also contribute to the development of intelligent systems. Toolkits like OpenAI Gym provide standardized environments for developing and benchmarking RL algorithms (Brockman et al., 2016).

While various commercial tools exist for data extraction, they often target enterprise users, auditors, or focus primarily on extraction rather than interpretation for laypersons. This study differentiates itself by specifically targeting non-financial individuals, leveraging the advanced contextual understanding and generative capabilities of OpenAI's platform to not only extract but also explain financial data from personal documents through a user-friendly interface. This focus on enhancing accessibility and financial literacy for a non-expert audience addresses a distinct gap in the current landscape of financial technology tools.

Research Gap

This study aims to address this gap by creating and implementing an AI-powered financial data extraction and interpretation system tailored specifically for non-financial individuals. The primary objective is to develop a user-friendly platform utilizing OCR, NLP (via the OpenAI platform), and ML technologies to automate the extraction, organization, and, crucially, the *comprehension* of financial information from personal documents. Specific objectives include identifying the needs and challenges of non-financial individuals in managing personal financial statements, developing an AI-powered platform using relevant technologies to automate the process, and evaluating the developed system's usability, accuracy, and effectiveness in empowering users with limited financial expertise.

Objective of the Study

The primary objective of this project is to create a system that automates the extraction, organisation, and comprehension of financial information from personal documents such as financial statements.

1. To identify non-financial individuals' needs and challenges in managing personal financial statements.
2. To develop an AI-powered platform utilising OCR, NLP, and machine learning technologies to automate the financial information.
3. To evaluate the developed system's usability, accuracy, and effectiveness in empowering individuals with limited financial expertise.

METHODOLOGY

This study employed the Waterfall system development model, a sequential approach involving distinct phases: Requirements Gathering, Design, Implementation, Testing, and Deployment, as shown in Figure 1. This structured methodology provided a systematic progression for developing the Financial Data Extraction system.

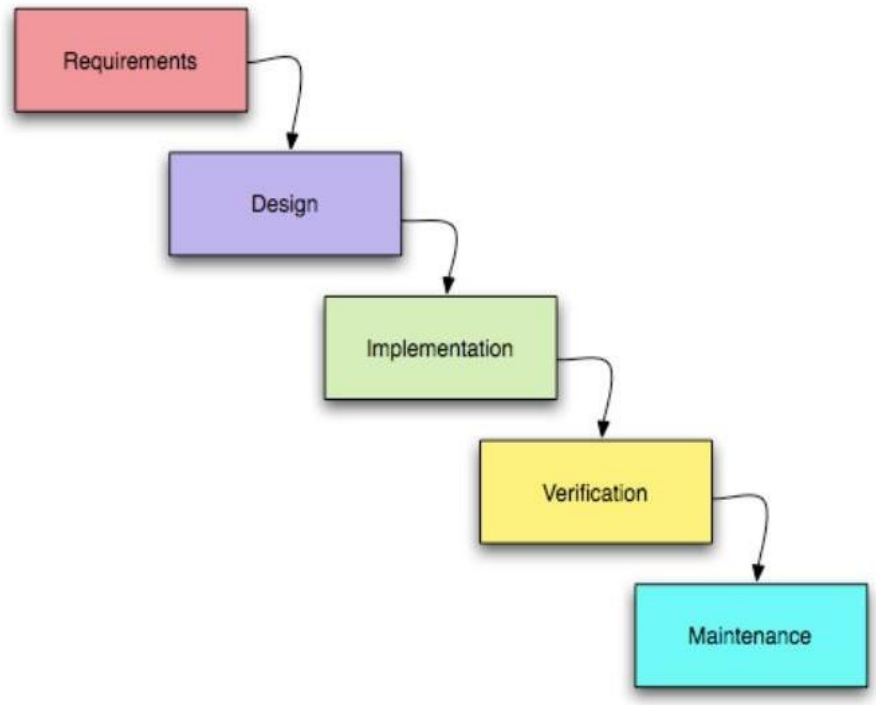

Figure 1 The waterfall system development model

Requirements Analysis

The initial phase focused on clearly understanding the needs and challenges of the target users—non-financial individuals managing personal financial documents. Stakeholder engagement involved distributing questionnaires to gather diverse perspectives. A structured questionnaire, mixing closed and open-ended questions, was developed and distributed to a sample of potential end-users represented by art students in the initial survey development phase, chosen for their likely non-financial background. Key findings from this initial requirement gathering indicated that most potential users were young adults, around 73.3% aged 18-20. Many expressed high confidence, with 80% rating 4 or 5 in understanding financial statements generally, yet identified specific challenges with balance sheets, financial ratios, and overall structure/terminology as shown in Figure 2.

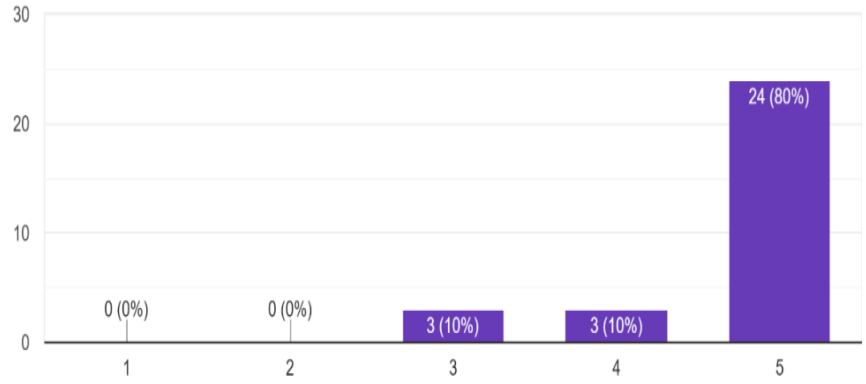

Figure 2 The understanding financial statements

100% reported not using any tools currently to aid understanding. The preferred learning method was written explanations and examples, with 70%, 13.3% prefer personalized feedback and guidance, 13.3% require visual aids, and lastly 3.3% choose interactive tutorials or simulation, as illustrated in Figure 3.

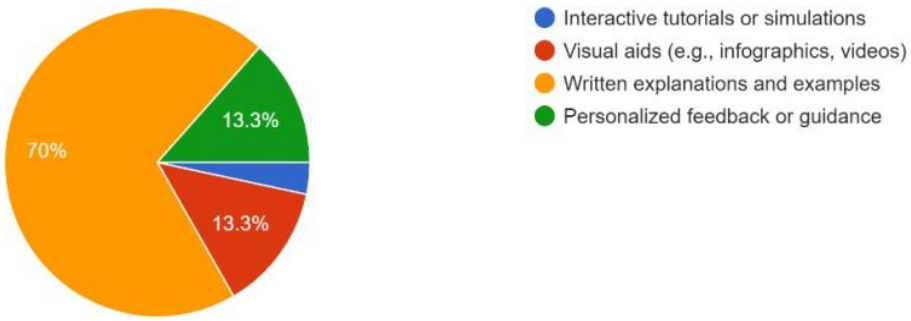

Figure 3 The preferred learning method

There was overwhelming interest (86.7% rating 4 or 5) in a system that simplifies and explains financial statements. These insights confirmed the need for an educational, user-friendly tool and guided the system's design as shown in Figure 4.

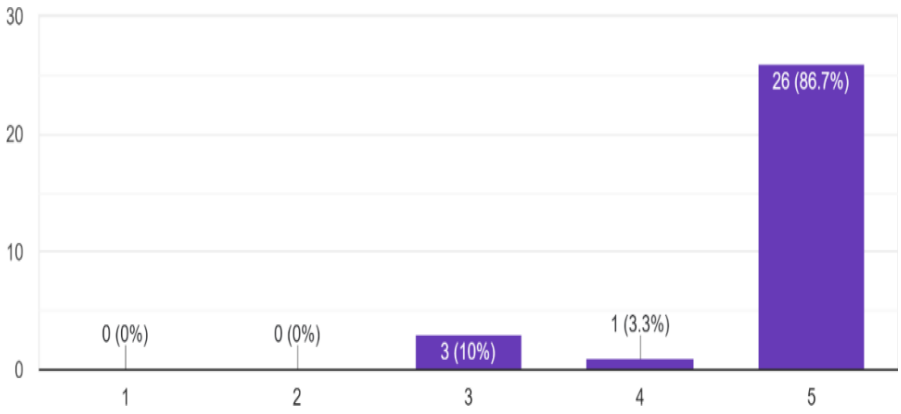

Figure 4 The need for an educational, user-friendly tool and guided the system's design

System Design

The system architecture facilitates the automated extraction and management of financial data from personal documents for non-financial individuals. The architecture comprises several key components: First, an upload module allows users to submit image files of financial statements. These files are processed using OCR

(Optical Character Recognition) technology to extract text accurately from images. Next, the extracted text undergoes natural language processing (NLP) through OpenAI, enabling the system to interpret and extract relevant financial information such as amounts and transaction details. The system includes a user-friendly web interface that displays OCR results and AI-generated insights, facilitating easy interaction. The core workflow involves the process as shown in Figure 5.

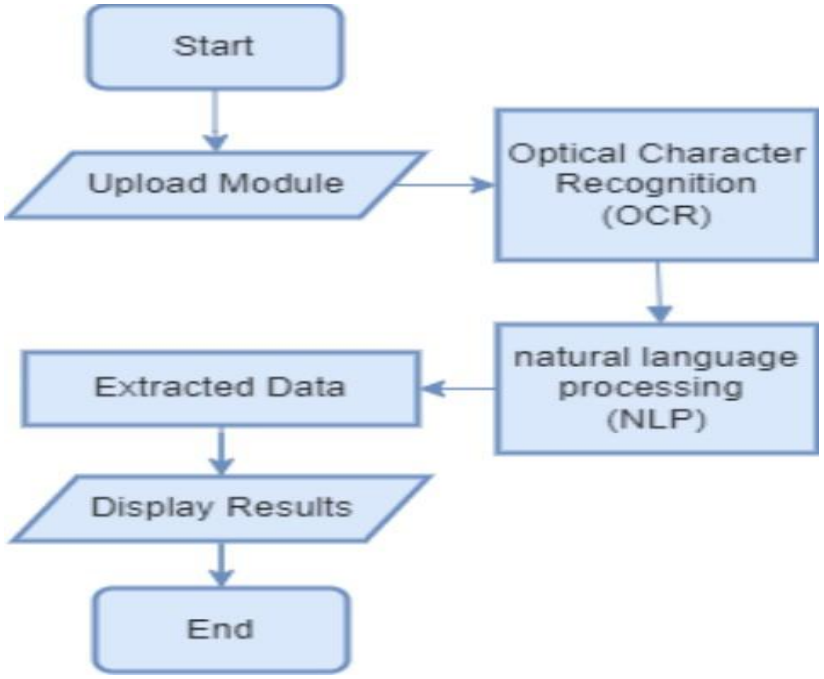

Figure 5 The core workflow

Algorithm specifications detailed the steps for preprocessing, OCR conversion, text cleaning, NLP application, data validation, and results display. Integration with the OpenAI API was a key design focus, leveraging its advanced language model capabilities. User Interface design prioritized intuitiveness and clarity, using a dark background with light text for readability. Scalability and performance were considered through efficient algorithm design and potential for cloud deployment. The algorithm is illustrated in Figure 6.

```
Algorithm ExtractAndManageFinancialData

Input: imageFile (Uploaded financial document as image)

Output: Displayed financial data and insights on web interface

BEGIN

imageFile ← GetUploadedFile()

preprocessedImage ← PreprocessImage(imageFile)

extractedText ← PerformOCR(preprocessedImage)

nlpResponse ← CallOpenAINLPModel(extractedText)

financialData ← ParseFinancialInformation(nlpResponse)

DisplayOnWebInterface(extractedText, financialData)

END
```

```
Function PreprocessImage(image):

Return cleanedImage

Function PerformOCR(image):

Return text

Function CallOpenAINLPModel(text):

Return response

Function ParseFinancialInformation(response):

Return structuredData

Function DisplayOnWebInterface(text, data):

Render(text)

Render(data)
```

Figure 6 The algorithm

Testing and Validation

A multi-faceted testing strategy was employed to ensure functionality, accuracy, and usability. Functionality Testing: Unit tests verified individual components like OCR accuracy and AI response generation. Integration testing ensured seamless interaction between modules. System testing validated overall behaviour against requirements. Specific test cases involving uploading valid/invalid file types and checking storage were executed. Data Sampling & Accuracy: Tests used real financial statements from IIUM students to evaluate the system's capability to accurately extract relevant information. Finance experts compared extracted data against originals to assess accuracy and interpretation. Test cases confirmed successful text extraction and AI response generation for sample data. Performance Testing: Load, stress, scalability, and endurance testing were planned to assess the system's reliability and efficiency under various conditions.

User Acceptance Testing (UAT): This crucial phase involved actual end-users of finance staff evaluating the system's suitability for non-experts interacting with the system to validate functionality, usability, and alignment with expectations using meticulously crafted test cases based on user stories. Feedback on ease of use, data reliability, and overall satisfaction was collected via a dedicated User Evaluation questionnaire.

RESULTS

System Functionality

The implemented system successfully integrated the designed components. Users could navigate the interface to upload financial document images. The OCR module accurately processed these images, extracting text which was then passed to OpenAI. The OpenAI module generated relevant responses based on the extracted financial data and predefined prompts. Functionality tests confirmed that file uploads handled valid image types such as JPG and PNG correctly and rejected invalid types likes PDF and TXT. OCR processing successfully extracted text from sample documents, and the OpenAI module provided accurate responses based on this text. Test case generation using specific sample data demonstrated successful end-to-end processing for file upload, OCR, and AI response generation. Figure 7 and 8 show the OCR processing screen and Open AI response screen.

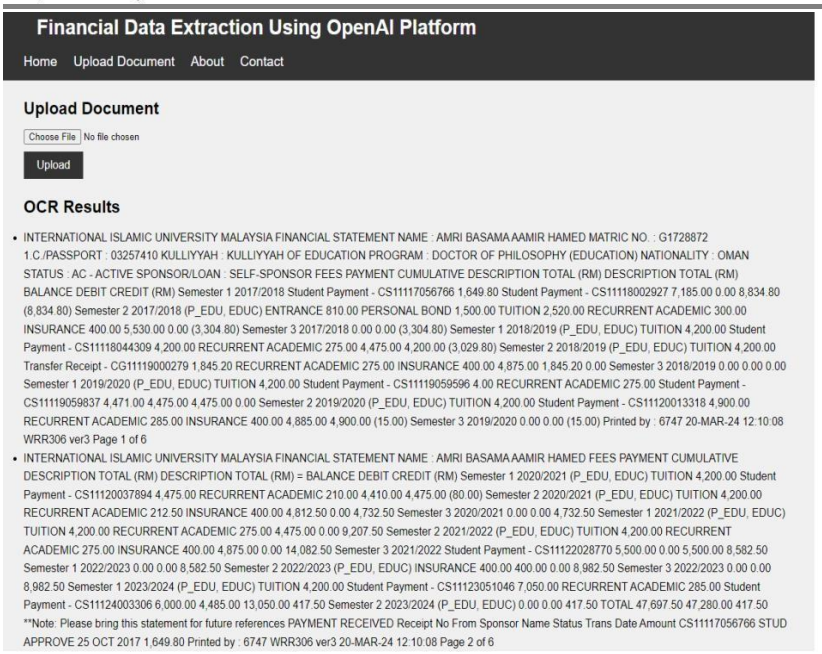

Figure 7 The OCR processing screen

Figure 8 The Open AI response screen

User Evaluation Findings

User Acceptance Testing (UAT) was conducted with finance staff providing feedback on the system's potential utility and effectiveness for non-financial users. Demographics showed participants were mostly female and predominantly over 40 years old. The evaluation yielded overwhelmingly positive results across various dimensions.

Effectiveness & Accuracy: 50% rated the system's effectiveness in accurately reading and extracting data as excellent, with another 33.3% rating it 4. Similarly, 50% rated its ability to interpret and understand context as excellent, with 16.7% rating it 3. Accuracy of insights was also rated highly, with 50% giving a score of 5 and 33.3% a score of 4.

Usability & Navigation: Ease of navigating and using the tools was rated highly, with 50% scoring 4 and 33.3% scoring 5. Interface intuitiveness and overall usability received similar high ratings of 83.3%.

Functionality & Satisfaction: Participants expressed high satisfaction with the system's ability to analyse trends and patterns, with 50% scoring 4 and 50% scoring 5. The system's handling of complex documents and diverse formats was rated very good with a 66.7% score of 4 and excellent with a 33.3% score of 5.

Efficiency & AI Impact: The system was perceived as significantly improving efficiency, with 50% rating its impact as excellent, scoring 5, and 16.7% as good, scoring 3. Critically, the impact of integrated AI and ML in enhancing the depth and speed of analysis was rated overwhelmingly positively, with 83.3% indicating a significant effect with a score of 5.

Overall Results Analysis

The results demonstrate the successful implementation and positive reception of the automated financial statement analysis system. The combination of OCR for accurate text capture and OpenAI's NLP for interpretation and insight generation proved effective. The consistently high ratings across accuracy, usability, and efficiency metrics from the user evaluation affirm the system's effectiveness in meeting user needs, particularly highlighting the significant positive impact attributed to the AI/ML integration. The overwhelmingly positive feedback suggests the system successfully streamlines complex tasks and holds strong potential for enhancing financial literacy and decision-making for its intended non-expert audience.

DISCUSSION

The project successfully achieved its primary objectives. Firstly, the requirements analysis phase, through questionnaires and stakeholder engagement simulations, effectively identified the core challenges faced by non-financial individuals, difficulty with financial jargon, complex structures, and a lack of accessible tools. Secondly, an AI-powered platform utilizing OCR and OpenAI's NLP was successfully developed, integrating these technologies into a cohesive system capable of automated extraction, organization, and comprehension of financial information from diverse statements. Thirdly, the evaluation phase incorporates functionality testing and user acceptance testing. It validated the system's high usability, accuracy, and effectiveness in empowering individuals with limited financial expertise.

The key finding is the system's demonstrated ability to bridge the gap between complex financial information and non-expert users. The integration of OpenAI's sophisticated NLP capabilities was perceived as highly impactful, with 83.3% rating 5, enabling not just data extraction but meaningful interpretation and simplification. This suggests that advanced AI models can significantly enhance accessibility and understanding, moving beyond basic data capture. The high usability scores indicate that the user-centric design approach was successful, creating an intuitive interface that facilitates engagement even for those unfamiliar with financial analysis tools. The implication is that such systems can serve as powerful educational tools, fostering greater financial literacy and enabling more confident personal finance management by demystifying complex documents.

Compared to existing commercial data extraction tools often focused on enterprise or auditing tasks like Nanonets and Data Snipper, this system carves a distinct niche by prioritizing interpretation and explanation for individual, non-financial users managing personal documents. Its core strength lies in the synergistic combination of OCR for data capture and advanced LLM via OpenAI for contextual understanding and user-friendly presentation. Limitations include the evaluation being performed by finance staff for non-expert use. The potential dependency on the OpenAI API's performance and cost structure, and the scope of document types tested during development. Further testing with a broader range of non-financial end-users and document formats would be beneficial.

CONCLUSION AND FUTURE RECOMMENDATIONS

This study successfully designed, developed, and evaluated an innovative system for financial data extraction and interpretation targeted at non-financial individuals. By integrating OCR technology with the advanced NLP capabilities of the OpenAI platform, the system effectively automates the process of reading, understanding, and explaining information from personal financial documents. Functionality tests confirmed the system's operational integrity, while user evaluations revealed exceptionally high levels of satisfaction regarding accuracy, usability, efficiency, and the transformative impact of the integrated AI. The project demonstrates the significant potential of leveraging advanced AI to democratize access to financial

information, enhance financial literacy, and empower individuals to manage their personal finances more effectively.

Based on the project's success and evaluation findings, several future recommendations emerge. Firstly, expanding language support and compatibility with a wider array of document formats would significantly broaden the system's accessibility and global usability. Secondly, integrating more sophisticated AI algorithms for predictive analytics could offer users proactive insights and forecasting capabilities, further enhancing financial planning. Thirdly, incorporating Natural Language Generation (NLG) could enable the automatic creation of comprehensive financial summaries in a highly readable, narrative format. Continuous refinement based on ongoing user feedback and iterative testing will be essential for adapting to evolving user needs and technological advancements. Finally, enhancing data security measures and ensuring compliance with privacy regulations remain paramount to maintain user trust as the system evolves. Pursuing these enhancements can position the platform as a leading tool in advancing personal financial management practices.

REFERENCES

1. Akanksha, E., Sharma, N., & Gulati, K. (2021, April). Review on reinforcement learning, research evolution and scope of application. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1416–1423). IEEE. https://doi.org/10.1109/ICCMC51019.2021.9418386
2. Boubaker, S., Gounopoulos, D., & Rjiba, H. (2018). Annual report readability and stock liquidity. CGN: Disclosure & Accounting Decisions (Topic).
3. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). OpenAI Gym. arXiv preprint arXiv:1606.01540. https://doi.org/10.48550/arXiv.1606.01540
4. Chakraborty, I., Leone, A. J., Minutti-Meza, M., & Phillips, M. K. (2019). Financial statement complexity and bank lending. S&P Global Market Intelligence Research Paper Series.
5. Chew, P. A., & Robinson, D. G. (2012). Automated account reconciliation using probabilistic and statistical techniques. International Journal of Accounting & Information Management, 20(4), 322–334. https://doi.org/10.1108/18347641211287763
6. Clementina, K., & Idume, G. (2015). Bank reconciliation statements, accountability and profitability of small business organisation. Research Journal of Finance and Accounting, 6(21), 21–30.
7. Costales, S. B. (1979). The guide to understanding financial statements.
8. Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In Machine learning techniques for multimedia: Case studies on organisation and retrieval (pp. 21–49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2
9. Fahad Mon, B., Wasfi, A., Hayajneh, M., Slim, A., & Abu Ali, N. (2023). Reinforcement learning in education: A literature review. Informatics, 10(3), 74. https://doi.org/10.3390/informatics10030074
10. Gibson, C. H., & Frishkoff, P. A. (1983). Financial statement analysis: Using financial accounting information: Test bank to accompany.
11. Gruetzemacher, R. (2022, April 19). The power of natural language processing. Harvard Business Review. https://hbr.org/2022/04/the-power-of-natural-language-processing
12. Guay, W. R., Samuels, D., & Taylor, D. (2016). Guiding through the fog: Financial statement complexity and voluntary disclosure. Research Methods & Methodology in Accounting eJournal.
13. Gupta, A., Dengre, V., Kheruwala, H. A., Raut, R. D., & Kamble, S. S. (2020). A comprehensive review of text-mining applications in finance. Finance Innovation, 6(39). https://doi.org/10.1186/s40854-020-00205-1
14. Higson, C. (2006). Financial statements: Economic analysis and interpretation.
15. Jensen, K. T. (2023). An introduction to reinforcement learning for neuroscience. arXiv preprint arXiv:2311.07315. https://doi.org/10.48550/arXiv.2311.07315
16. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. Multimedia Tools and Applications, 82(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4
17. Madhavi, A., & Sreedivya, B. (2017). A novel bank statements reconciliation using message transfer parser. In 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1–6). IEEE. https://doi.org/10.1109/ICCIC.2017.8528707

18. Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE Access, 8, 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542
19. Mourik, C. V., & Walton, P. (2013). The Routledge companion to accounting, reporting, and regulation. Routledge.
20. Rakshit, A., Mehta, S., & Dasgupta, A. (2023, June). A novel pipeline for improving optical character recognition through post-processing using natural language processing. In 2023 IEEE Guwahati Subsection Conference (GCON) (pp. 1–6). IEEE. https://doi.org/10.1109/GCON57805.2023.10236198
21. Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. In K. Shaalan, A. Hassanien, & F. Tolba (Eds.), Intelligent natural language processing: Trends and applications (Vol. 740, pp. 373–397). Springer. https://doi.org/10.1007/978-3-319-67056-0_18
22. Sherman, H. D., & Young, S. D. (2016). Where financial reporting still falls short. Harvard Business Review, 94, 17.
23. Tshehla, M. (2022). An investigation of the impact of material misstatements on the quality of financial reporting for a public sector. In Proceedings of the 5th International Conference on Business, Management and Finance.
24. Veal, L. (2005). Tax knowledge for undergraduate accounting majors: Conceptual v. technical.
25. Xing, X. (2021). Financial big data reconciliation method. In 2021 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE) (pp. 260–263). IEEE. https://doi.org/10.1109/ISAIEE53056.2021.00065
26. Yenduri, G., Srivastava, G., Maddikunta, P. K. R., Jhaveri, R. H., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. arXiv preprint arXiv:2305.10435. https://doi.org/10.48550/arXiv.2305.10435