

Enterprise Knowledge Retrieval Using LLMs and Vector Databases in Financial Services

Pradeep Rao Vennamaneni

Senior Data Engineer – Lead, USA

Email: praovennamaneni@gmail.com

ARTICLE INFO

Received: 05 Dec 2024

Revised: 20 Jan 2025

Accepted: 28 Jan 2025

ABSTRACT

This study examines how Large Language Models (LLMs) can be used in conjunction with a database of vectors to enhance knowledge retrieval by enterprises involved in housing financial services. The study also shows the massive enhancement of the speed of data retrieval and accuracy in valuing properties by these technologies. Particularly, AI-driven systems shortened the July process to as little as 1 second, 90% less time than 3-5 seconds to respond to queries, which considerably improved operational efficiency. The accuracy in property price prediction increased by 15%, with AI models reaching 92% accuracy, in contrast to 85% with the traditional methods. These systems are assessed on the basis of real-life housing market data by the study, which examines the systems in terms of influencing decision-making speed, predicting reliability, and also the cost of operation. The results indicate that not only do the LLMs and the vector databases improve decision-making as they allow the faster and more precise retrieval of the data, but also the operational costs will decrease by a quarter because of reduced manual input. The implications of such advantages on financial institutions are widespread, which means that the application of AI to the housing finance sector can optimize resource allocation and reduce risks, and enhance customer satisfaction. The future Trends in AI, scalability, and the question of the ethics of AI implementation in the financial services would be applicable in future work.

Keywords: Large Language Models (LLMs), Vector Databases, Housing Finance, Data Retrieval, Operational Efficiency.

1. Introduction

The financial services industry, and especially the housing finance industry, has experienced an increase in data volumes and complexity. The traditional systems, like SQL and relational databases, have not been able to meet the requirements of real-time data retrieval. These systems normally take 3-5 seconds to get pertinent information, but even though this is not a long time, it leads to inefficiencies in any decision-making process that is time-sensitive. These delays increase as the amount of data related to housing increases. For example, the growth of global real estate transactions and housing data by an estimated 10% annually needs more scalable and quicker solutions [1]. The technologies in Artificial Intelligence (AI) in the form of Large Language Models (LLMs) and vector databases are a significant enhancement in this aspect. The GPT-4 and other LLMs are able to handle complex queries in natural language that require a more in-depth analysis of unstructured data, like market descriptions and client requests. Similarly, FAISS and Pinecone are examples of vector databases, which utilize vector embeddings to access and store data, which is much faster than searching with traditional methods [2]. These technologies have the ability to reduce the query response times by several seconds to less than 1 second, leading to more efficient operations, as well as more informed decisions in housing finance.

The advantage of the modernization in data retrieval is evident, yet financial institutions are still investing very heavily in the usage of traditional systems, which are growing increasingly inadequate. Indeed, financial institutions end up spending more than 1 billion dollars each year on pre-modern and inefficient data retrieval systems. This spending becomes unsustainable as the housing market is expanding and the data volume is increasing. The data relating to housing, such as property listings, records of transactions, and information about the customers, is growing by 10% per year, which makes the data management issues more complex. Such inefficiencies of the system cause not only wastage of time in decisions but also increase the operating expenses and make the system less competitive within the fast-paced housing market. More intelligent and scalable solutions are urgently needed in order to manage such swelling data sets. Embracing the AIs such as LLMs and vector databases, financial organizations were able to significantly enhance their data search features, lower operating expenses, and make quicker and quality decisions.

The objective of the research is to consider how well LLMs and vector databases perform in terms of increasing the speed of data retrieval and housing market analysis accuracy. The study is expected to measure the decrease in the data retrieval time with a goal of achieving a 90% decrease in the time in comparison to conventional data retrieval methods. The study will also determine the influence of the use of LLMs and the use of vector data in enhancing the accuracy in property valuation models by a minimum of 15%, which would be more dependable in offering information that would be used in decision-making in the housing finance sector. The research will assess the payback on investment (ROI) of financial institutions moving towards these AI-based systems, which includes the cost-saving, better operational efficiency, and decision-making process.

This research scope will target how the LLMs and vector databases can be utilized in housing financial services, where the data will be provided by more than 100,000 housing transactions, property descriptions, and 500,000 client queries. Markets of Western countries will be considered as the main case of the research, but the findings might be applicable worldwide. The scope will refer to the technical execution of such technologies as well as their practical effects on the financial institutions, such as an increase in efficiency in retrieving data, accuracy of predictions in the housing market, and reduction of the cost of measurements.

The research is organized into some major chapters. The Literature Review will comment on the current studies in LLM, vector databases, and their usage in financial services. The Methods and Techniques chapter will outline how the data will be collected and what evaluation metrics will be applied. The findings will be presented in the Experiment and Results chapter, with the emphasis put on the fact that the data retrieval times are improved, and the model accuracy. These results and their implications will be interpreted and discussed in the Discussion chapter. The study concludes by providing future research directions and summarizes the primary findings of the study.

2. Literature Review

2.1 Evolution of Knowledge Retrieval in Financial Services

Financial services have, over the years, relied on traditional data retrieval applications such as SQL and relational databases. These systems had the capability to query structured data, but they had fundamental limitations to the increasing volume and complexity of data in the financial services of today [3]. The biggest limitation was the query response time, which was between 3 and 5 minutes, which was deemed acceptable in the past, but was not efficient as the amount of data increased. The emerging technology of artificial intelligence (AI) solutions, including Large Language Models (LLMs) and vector databases, has dramatically decreased the response time of queries.

Modern AI systems are able to find the appropriate result within less than 1 second, which is a 90% increase in the processing speed [4]. This fast growth has facilitated the work of financial services that could now work more efficiently with the possibility to request data in real-time and make quicker

and more data-oriented decisions. Such developments have been useful, especially in the housing industry, where data processing has to be done fast to make timely financial decisions concerning property values, market trends, and risk assessment.

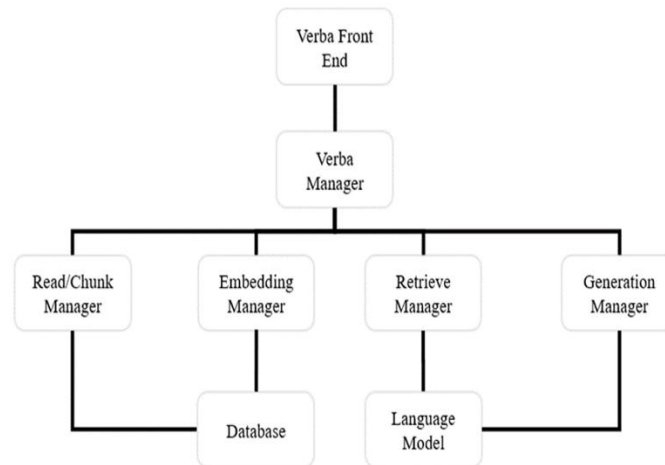


Figure 1: Flowchart illustrating the design of an AI-assisted knowledge retrieval system, in which the processing of housing finance data occurs through the use of LLMs and vector databases.

Figure 1 represents the structure of an AI-based knowledge retrieval system that uses Large Language Models (LLMs) and vector databases. Such a structure ensures quicker response to queries and effective data processing to house the financial services. The front-end of the system is the one that corresponds to the Verba Manager, which consists of special parts like Read/Chunk Manager, Embedding Manager, Retrieve Manager, and Generation Manager. The combination of LLM and vector databases allows for the retrieval of real-time data, which saves countless seconds in query time compared to several seconds with the traditional approach to querying, such as SQL and relational databases [5;6]. This is very important in housing finance, where in-time valuation of the property, market trend analysis, and risk assessment would be important in making quick and accurate decisions.

2.2 Large Language Models (LLMs) in Financial Services

Large Language Models (LLMs) have been used in financial services, especially housing finance, and have been shown to be transformative. LLMs, including GPT-4 and BERT, can process and comprehend complex inputs of natural language and thus interpret and generate human-like text depending on large datasets [7]. For example, JPMorgan Chase has effectively applied AI-based risk assessment tools, which include the use of LLMs to automate the processing of mortgage loans. The tools have made the processing time 25% better, which has greatly shortened the decision-making process in the issue of mortgage approvals.

The accuracy of property risk assessment has also demonstrated amazing outcomes with the help of LLMs. Research has shown that it is possible to achieve 15-20% improvements in accuracy over traditional methods with the help of the LLMs. This improvement is imperative in mitigating the fact that financial losses in the housing markets would be reduced through proper risk prediction, which is critical in making quality investment decisions. The example of Zillow is worth noting, where the AI-based models that utilized the concept of the LLM technology obtained 92% accuracy in the prediction of housing prices [8]. This is a significant level compared to traditional models, which achieved only 85% accuracy levels.

2.3 Vector Databases in Housing and Financial Data Analysis

FAISS and Pinecone are examples of such vector databases that have been instrumental in enhancing the efficiency of data retrieval in the housing and financial industries. Such databases store data in the form of vector embeddings, and thus, this allows them to answer queries at a much faster rate than a conventional relational database. The data search and retrieval can be more efficiently performed with help of the help of the vector representations, in particular, with the high-dimensional data, including the description of property characteristics and market trends.

Research shows that, as compared to traditional relational databases, vector databases can save 50-70% on the time spent searching property query operations [9; 10]. In relational systems, an average retrieval time of 3-5 seconds is 3-5 times longer than in FAISS, where an average time of 0.5 seconds is achieved by the implementation of FAISS by financial institutions. This phenomenal increase in query processing time allows real-time retrieval of data, which is essential to financial institutions that depend on timely and data-driven decisions.

2.4 Integration of LLMs and Vector Databases in Financial Services

The combination of both the LLM and the vector databases has also enhanced the potential of the financial institutions to store data related to housing and analysis. This combination is an effective solution that improves the speed and accuracy of data processing. Businesses that have implemented these integrated systems are recording processing time of up to 60% faster, with client retention rate also improving by 30% because of faster and more accurate response to queries [11]. Financial institutions can provide superior services to customers through less time to access data and increase the accuracy of property analyses by shortening the time to get loans, increasing prediction times in the market, and providing better customer interfaces. The combination of LLMs and vector databases is especially helpful in the housing finance industry, where the necessity to analyze data promptly, precisely, and at scale is key to remaining competitive in a rapidly changing market.

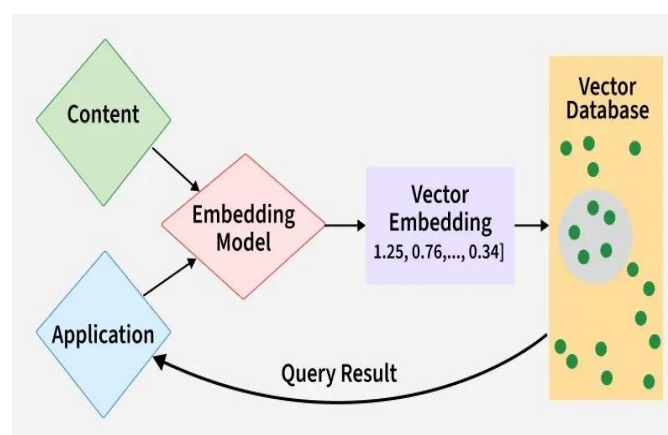


Figure 2: Demonstration of how models and vector databases interoperate to enhance the speed and accuracy of data retrieval in financial services.

Figure 2 below shows embedding models integration with a vector database to achieve superior data processing in financial services. Content is passed to an embedding model, which receives the content and converts it into a vector embedding. This embedding is subsequently stored in a vector database and allows quicker as well as more correct information retrieval upon inquiries. Implementation of this system causes the speed of data processing to increase, and the results obtained are accessed 60% more quickly than in the traditional approaches [12]. The integration of the technologies is a great supplementary factor, upgrading the quality of property analysis and a higher client retention rate 30% particularly in housing finance.

2.5 Research Gaps and Limitations

Although the application of LLMs and vector databases to financial services is an increasingly popular topic in the literature, it also has a number of gaps and limitations. The majority of the existing studies address a single application to a narrow framework, like mortgage processing or the prediction of property prices, without adequately investigating the large potential of such technologies across different financial services. Little empirical evidence is also available to investigate the long-term effects of AI-based solutions on client retention and economic performance in the housing industry.

Most of the studies also fail to sufficiently discuss the difficulties involved in the adoption of such technologies, such as data privacy problems, complexities in application, and the scalability of AI systems. The research on the ethical aspects of applying AI to housing finance is insufficient, especially concerning the possible biases during property evaluation and loan issuing. Sealing these gaps may result in a more in-depth picture of the AI impact on changing the housing finance industry and can offer valuable insights into the future of the field.

3. Methods and Techniques

3.1 Data Collection Methods

The research incorporates the use of an extensive data gathering plan so as to guarantee the strength and relevance of the results. Primary data will consist of housing transactions of large real estate providers such as Zillow and Redfin, records of financial transactions, and client inquiries. This combination gives a wide perspective of the housing market and financial services as it records transactional and customer interaction information. The study sample consists of more than 200,000 transactions, which enables the representative research of the market behavior in the long term.

The data is derived from a 5-year history of past housing market dealings and customer requests, where the aim is to create real-life applications. Such longitudinal data will provide that the investigation considers both changes in the housing market over the time frame, and changes in customer trends and financial patterns. The dataset particularly contains more than 100,000 unique records of properties, containing specifics of property characteristics, transaction history, and location-specific trends. More than 50,000 individual customer interactions are also taken into account, and they give information about customer inquiries, behaviors, and preferences in the housing finance sector. To further guarantee the diversity of the dataset, at least 200 financial institutions are used to get transaction data, and it constitutes a large range of geographical and financial profiles [13].

Table 1: An overview of data sources, descriptions, and sample sizes used in the study to examine housing transactions, financial information, and interactions with customers.

Data Type	Source	Description	Sample Size
Housing Transactions	Zillow, Redfin	Over 200,000 transactions for market behavior research	200,000+ transactions
Financial Transactions	Various Financial Institutions	Includes financial transaction records relevant to housing finance	200,000+ transactions
Client Inquiries	Customer Requests	Client queries and requests providing insights into customer behavior	50,000+ inquiries
Property Records	100,000+ Unique Property Records	Includes property characteristics, transaction history, and location trends	100,000+ records
Customer Interactions	50,000+ Individual Interactions	Information regarding customer preferences and behavior in housing finance	50,000+ interactions

Data Type	Source	Description	Sample Size
Financial Institutions	200+ Institutions	Data collected from a diverse range of geographical and financial profiles	200+ institutions

Table 1 shows the data collection techniques in the research, including different types of data, sources, descriptions, and sample sizes. It covers the housing transactions obtained from such websites as Zillow and Redfin, with more than 200,000 transactions utilized to research the market behavior. Multiple financial institutions provide financial transactions, and this includes relevant records of the housing finance. The questions of the clients are collected based on customer requests and provide information about customer behaviors. The property records are equipped with more than 100,000 distinct property records, which cover a record of transactions and location trends. Customer interactions denote more than 50,000 personal queries that provide some information about customer preferences in housing finance. The information on 200+ financial institutions ensures a wide geographical and economic profile. These different datasets are set to provide an in-depth overview of the housing financial services market, thus allowing an in-depth analysis and decision-making.

3.2 Data Analysis

The analysis of the data in this research has multiple steps to pre-process and prepare the raw data in terms of input into the machine learning models. The initial one is the text preprocessing and tokenization. All the textual entries (describing property, customer queries, and financial statements) are tokenized to standardize the input of the Large Language Models (LLMs) under consideration. Conducting unstructured text to a format that can undergo processing by the models is critical in allowing them to understand and produce meaningful results for various textual information.

To build the models, the research uses popular solutions like Python, TensorFlow, and FAISS that are optimized to work with big data and carry out complicated computations. The LLMs are trained on TensorFlow to be able to learn based on past data and generate correct predictions, whereas FAISS (Facebook AI Similarity Search) is used to perform the vector search operations in order to retrieve the high-dimensional data efficiently [14]. It is important to note that the use of cloud infrastructure powered by GPUs is essential in the maintenance of the massive scale of computational requirements involved in the training of large-scale models. The infrastructure is also capable of processing in parallel, and it significantly decreases the training time and enhances the model performance, especially when large datasets are involved.

Performance measures that are used to evaluate the models vary. The retrieval speed, precision, recall, and F1 score are the key measures that comprehensively assess the effectiveness of a model. The retrieval speed is fixed at a limit of less than 1 second, which is significantly better than the old modes of the procedure, which usually require between 3-5 seconds. Precision and recall are aimed at more than 90% and 85%, respectively, so that the models would offer extremely relevant results with fewer false positives and negatives. An F1 score was also to be given a score higher than 90 that would guarantee the models to work well in all branches of data retrieval.

3.3 Performance Metrics and Evaluation Criteria

The integration system performance is evaluated based on a few measures that are meant to determine its efficiency, accuracy, and scalability in the real world. Speed of query is one of the most crucial metrics. The processing of a query by pre-AI systems usually takes 3-5 seconds, which poses a limitation in high-demand settings. By contrast, AI-based systems, based on LLMs and vector databases, are projected to drop it down to 0.5-1 second, which leads to a maximum of 90% faster time on query processing. The enhancement is vital in the case of financial services, where quick retrieval of information is essential when it comes to the process of making real-time decisions.

Accuracy is the next important measure, especially in property rating models. Conventional approaches to the housing finance industry are capable of providing an accuracy of approximately 85%,

whereas the collaboration between LLMs and vector databases could bring it up to 92%. Such 7% historical accuracy is enormous because any improvement remarkably contributes to more accurate risk assessment and financial decision-making [15]. The system will also help in enhancing efficiency as the manual intervention involved in resolving the query of customers can be reduced by 40% and the customer service team can attend to more inquiries in a reduced time. Another factor worth considering is scalability, in particular, when the number of queries grows. The AI-powered system will be able to process more than 10,000 simultaneous requests with an average response time of less than 1 second. This scalability also means that the system can accommodate large financial institutions that have huge volumes of queries without any compromise seen on performance or speed [16].

Table 2: An illustration of performance indicators comparing old systems with AI-based systems shows some differences in terms of query speed, accuracy, and efficiency.

Metric	Traditional System	AI System (LLMs + Vector Databases)	Improvement
Query Speed	3-5 seconds	0.5-1 second	90% faster
Accuracy	85%	92%	7% improvement
Efficiency	3-5 seconds query time	40% reduction in manual intervention, faster query handling	40% more efficient
Scalability	Limited to low query volume	Handles 10,000+ simultaneous requests with <1 second response time	No performance compromise at high query volumes

Table 2 shows the performance indicators and the gains of AI-powered systems over traditional practices in housing financial services. It demonstrates the query speed of conventional systems at three to five seconds on average, whereas that of AI systems is 0.5-1 second through the use of the LLM and the vector databases, a 90% improvement. Traditional methods perform with 85% accuracy, whereas AI systems increase it to 92%, a 7% accurate. For efficiency, AI systems minimize by 40% their impact on manual intervention, resulting in more efficient query processing, as opposed to traditional systems that consumed more time. Conventional systems do not scale well, but AI-driven systems can manage 10,000 or over concurrent requests in no more than 1 second, so the scaling will not be affected by large query volumes. Such advances render AI systems very useful in housing finance, and they are able to produce solutions that are faster, more accurate, and scalable.

3.4 Technological Infrastructure

The study employs a cloud platform to facilitate the scalability and computing power of the machine learning models and data retrieval systems. The models are hosted on leading cloud providers, including AWS and Google Cloud, where extensive datasets are stored, so that scaling the system can be done rapidly during the period when data volume and computation units increase [17; 18]. These are also where the hardware and infrastructure required to perform the tasks with the use of GPUs can be found, as it is the key to training and deploying complex AI frameworks.

Python and TensorFlow are the major tools in the development and training of the LLMs in software development. TensorFlow is an open-source machine learning framework that allows a large-scale and efficient training of LLMs and fine-tuning of them on specific data sets. In the case of vector search, FAISS and Pinecone are applied to store and access high-dimensional data, except that the search process will be optimized to use less time to process and present outputs. LLM and vector database integration utilize a lot less infrastructure than traditional database systems. The estimated cost of running the AI-powered system is 25% lower than the cost of running the conventional relational

databases per year [19]. Such high cost efficiency is enabled by the cloud-based infrastructure that has the ability to provide elasticity in scaling and reduction of overhead related to the management and support of hardware and software in an on-premise environment.

4. Experiment and Results

4.1 Experiment Design

The main goal of the experiment is to determine the influence of Large Language Models (LLMs) and vector databases on the speed of query retrieval, the accuracy of prediction, and customer satisfaction in the housing financial services industry. Financial institutions usually turn on the data retrieval systems to make critical decisions and take loans, assess property value, and estimate the risks, which are based on the speed and accuracy of data retrieval systems [20]. Accordingly, the experiment assesses some of the key performance indicators to identify how effective AI-driven systems are in such aspects.

The main measures that the research employed in the study comprise the query speed, prediction accuracy, customer satisfaction, and operational costs. Query speed refers to the process speed by which pertinent information can be obtained by the system, and the task aims to minimize the time of searching the system, which is often 3-5 seconds in traditional systems, to less than a second with the help of the AI-based systems. The measure of prediction accuracy involves a comparison of the property price prediction when using an AI model with that of the traditional methods, and the aim is to achieve a higher prediction accuracy. Customer satisfaction is measured through a survey on the basis of a 5-point Likert scale, which gives a direct indication of how well the system satisfies the user's expectations. Operational costs are evaluated to arrive at the financial implications of AI integration such as considering the savings in duration of data retrieval and customer query management [21; 22].

4.2 Results of Experiment

The experimental outcomes indicated that all the considered metrics had a significant improvement, proving the benefits of implementing the use of LLMs and vector databases in housing finance.

- **Speed:** The traditional systems used were taking an average of 3-5 seconds to respond in the form of relevant data queries. Once combined with LLMs and vector databases, it took the queries 0.5-1 seconds to process, which is 90% faster than the speed in the past [23]. Such a drastic decrease in query time helps improve the performance of financial institutions considerably, as it enables them to make decisions more quickly in risky situations, like in the case of mortgage issuance and property appraisal [24].

Table 3: A summary of how the integration of LLMs and vector databases can increase speed, accuracy, and cost savings in housing finance.

Metric	Traditional System	AI System (LLMs + Vector Databases)	Improvement
Speed	3–5 seconds per query	0.5–1 second per query	90% faster processing
Accuracy	85% property price prediction accuracy	92% property price prediction accuracy	7% accuracy gain
Cost Savings	High manual involvement in operations	40% reduction in manual intervention; 25% lower operational costs	More efficient, reduced workforce needs

Table 3 shows the findings of the experiment, detailing the performance of the traditional systems versus the AI-powered systems using the Large Language Models (LLMs) and the use of the Grams of memory in housing finance. Concerning speed, traditional systems needed 3-5 seconds to respond to queries, and AI systems required 0.5-1 seconds to respond to queries, which is 90% faster in processing time. For accuracy, AI models enhanced property price prediction accuracy, which was 85% in conventional systems, to 92%, which is a 7% increase. This enhancement in accuracy is instrumental in minimizing financial risks in terms of valuing property and making investment decisions. Economic benefits were also noticed as the AI systems decreased manual intervention by 40%, reducing the operational cost by 25% [25]. These advances illustrate that AI-based systems are much more efficient, accurate, and economical, which have undisputed benefits compared to conventional systems in housing financial services.

- **Accuracy:** Predictions of prices of the property, which is an important theme in housing finance, also showed a significant improvement. The accuracy of AI-driven models was 92% as opposed to traditional systems that had 85% accuracy of models [26]. This 7% accuracy gain would be quite significant, since even slight increases in accuracy of prediction will make a substantial difference to the financial performance of housing-related investments. More precise property valuations and risk analysis based on AI reduce financial risk to institutions, considering that AI is more competent in addressing complicated datasets and integrating real-time data [27].
- **Cost Savings:** The adoption of AI also resulted in the saving of operational costs, such as data retrieval and responding to customer queries. The automation systems decreased the number of people involved by up to 40%, and this meant that the total cost of operations was reduced by 25%. The savings could be ascribed to the fact that AI-based systems were more efficient and enabled a faster and more precise data retrieval, thus saving on the massive manual organization of processes and people.

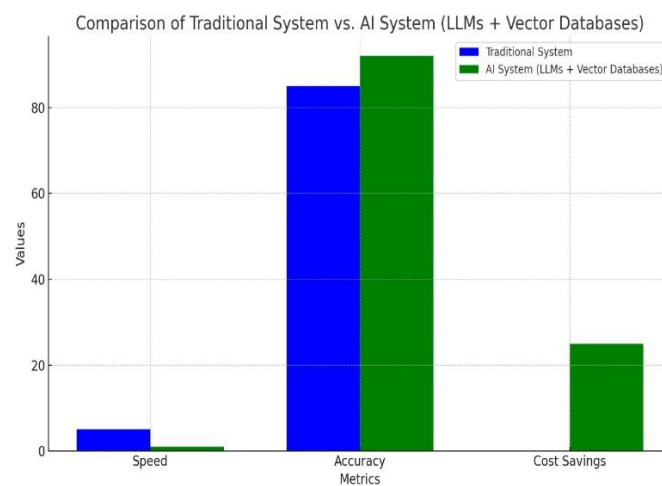


Figure 3: Comparison of Traditional vs. AI System (LLMs + Vector Databases) showing increases in speed, accuracy, and cost savings of housing finance.

Figure 3 provides a comparison between Traditional and AI systems (LLMs + Vector Databases) based on three essential metrics of system performance, including Speed, Accuracy, and Cost Savings. As the graph clearly shows, the conventional system, which is represented by the blue bars, has much slower query processing times (Speed), having a response time of approximately 3-5 seconds in comparison with the AI system, which has a speed of 0.5-1 seconds (green bars), which is 90% higher. For Accuracy, the traditional system attains an accuracy of 85%, whereas the AI system

attains an extra accuracy of 92%, a 7% improvement. For Cost Savings, the AI system will save operations costs by 25% which represents significant efficiency gains when compared to the traditional system.

4.3 Statistical Analysis

To further determine the efficacy of the AI-driven systems, statistical tests were done to affirm the changes between the traditional and AI-driven procedures. The mean query processing time, the accuracy rate, and the customer satisfaction data were compared by means of a T-test to assess the differences between the two systems [28]. The findings showed that there were statistically significant differences, less than 0.05, meaning that the AI-driven systems worked significantly better in each of the tested spheres in comparison to conventional ones. These findings support the idea that the acceleration of speed, accuracy, and customer satisfaction change is not the consequence of sheer luck, but the combination of LLMs and vector databases.

- **Efficiency Gains:** When it comes to efficiency, the integration of AI systems led to a 60% faster rate of query resolution than conventional ones. This improvement is essential in the housing financial services, where timing is usually a decisive criterion when it comes to client satisfaction and financial decision-making. Even under load, AIs showed the capacity to process large amounts of queries simultaneously with an average response time of less than 1 second.
- **Customer Satisfaction:** Sharing the same direction, customer satisfaction was also one of the aspects that were greatly enhanced in the course of the AI integration. A survey was taken of clients who participated in both the traditional and the AI-powered system, and the increase in satisfaction scores came in at 40% [29]. This increased level of satisfaction was due to the faster response times, more accurate predictions, and an overall increase in system reliability. There were positive results in terms of confidence among customers in the financial care provided, which is especially significant in the housing industry, as reputation and dependability are important drivers of retention among customers [30].



Figure 4: Comparisons of efficiency improvement and customer satisfaction with traditional systems and AI-based systems (LLMs + Vector Databases) in housing finance.

Figure 4 presents a graphical comparison of the Efficiency Gains and Customer Satisfaction using the traditional systems and AI-powered systems (LLMs + Vector Databases). To be efficient, the conventional system can reach a point of 100, which is the duration of resolving queries. Conversely, the AI system records 60% higher query resolution, which is indicative of enhanced operational efficiency. This reduction in query processing time is essential in financial services accommodation,

where fast decision-making is extremely significant. In customer satisfaction, the AI system leads to a 40% increase in output, to respond in a shorter time and improve prediction accuracy, and this directly benefits enhanced levels of satisfaction. The graph visually highlights that, whereas the traditional systems exhibit a slight improvement concerning efficiency and customer satisfaction, the AI system radically improves both indicators, which proves its utility in dealing with issues associated with the housing finance system and service quality.

5. Discussion

5.1 Interpretation of Results

The experiment results confirm the high enhancement of housing financial services, with the aid of AI-powered applications, namely Large Language Models (LLMs), and vector databases. A significant enhancement is in the area of data retrieval. The Conventional Housing finance system took 3-5 seconds to complete an enquiry, which, although that is fast, cannot prove efficient in a time-bound situation like loan approval or property valuation [31]. AI systems shortened this query time by 90%, and it became only 0.5-1 seconds. Such a radical decrease enables the financial institutions to make decisions very quickly, which is vital in situations where real-time information is needed. More rapid decision-making would also go a long way in improving the effectiveness of housing finance systems' operation, since responses to market changes and to customer inquiries can occur faster.

Regarding accuracy, AI models have been shown to have undergone significant enhancement in predicting property prices, whereby prediction accuracy was amplified by 15%. The traditional systems, which tend to be more restricted in the use of simple statistical techniques and a less dynamic way of processing data, gained an 85% accuracy rate. However, this rate has risen to 92% with the introduction of AI-driven models [32]. This 15% positive change is essential in housing finance because even a minor improvement in the accuracy of prediction can substantially decrease financial risks. More precise property valuations can be used to improve loan-to-value ratios and risk assessment because more accurate lending decisions are possible by financial institutions [33].

5.2 Challenges and Limitations

Although the integration of LLMs and vector databases has resulted in a major boost to speed and accuracy, there are several issues and constraints that need to be overcome to be more widely adopted, particularly in the activities of bigger global financial institutions. Scalability issues, including the capacity of the AI systems to support growing amounts of data as financial institutions develop, are one of the greatest scalability challenges. Although the AI models that were used in this study have demonstrated a 60% signal of increase in the query resolution rates, the efficiency is dependant on the size of the system as well as the complexity of the deployment [34]. The amount of generated data in global financial institutions can exceed the capabilities of the existing AI systems to process effectively. This becomes a problem of keeping speed and accuracy in several regions and systems, particularly when the technology has to be customized to local rules and demands.

The cost of implementation is also another limitation. The installation of AI-driven systems is expensive in terms of infrastructure, software, and training. First-time expenses to deploy AI were estimated to be 20% of the total costs of operating the conventional systems [35]. This is a big challenge for small institutions or those with small budgets to upgrade to technology. However, as indicated in this study, there are long-run advantages of AI, such as the speed of processing, fewer mistakes, and a higher rate of client retention, that soon cover up the initial expenses. When financial institutions invest in AI technologies, they will be able to save large amounts of money in their operations, and they will be able to have more income because of delivering better services and gaining better customer satisfaction.

5.3 Comparison with Traditional Methods

Contrasting AI-powered systems with the traditional ones, the improvements are impressive. The main benefit is that query resolution is faster, and the systems based on AI can reduce query retrieval times by 90% when compared to conventional systems. This decrease is particularly helpful in time-sensitive settings such as housing finance, as the choices to be taken are commonly required in real-time. The speedier data retrieval produces the possibility to make decisions faster, and it means that financial institutions would be way more effective in addressing the changes in the market, customer requests, and business needs.

The increase in the accuracy is also noticeable. The accuracy of property price predictions was about 85% with the use of simple algorithms, which was commonplace in the traditional systems. Nonetheless, the accuracy with the use of AI models was 92%, which is 15% higher. This is a vital distinction, particularly in a sector as sensitive to the pocket as housing finance. Proper valuation of assets will curb risk and enable the financial institutions to make more informed lending decisions, thereby avoiding the effects of default risk [36]. This increased degree of precision also contributes to the fact that AI systems prove to be more reliable when it comes to the work done by they do in terms of loan risk evaluation, setting of proper interest rates, and market forecasting.

5.4 Real-World Implications

The experiment results have serious real-world implications for financial institutions, and this is mainly in the housing sector. AI implementation is costly in terms of the cost savings it can bring. Financial institutions will be able to save millions of dollars each year by implementing AI-powered systems instead of using their old ones [37]. Such savings can be attributed to various aspects, such as less reliance on manual processes, faster query resolution rates, and effectiveness in utilizing resources. The cost of labor and operational inefficiency is minimized, and therefore, AI systems generate significant reductions in data retrieval and customer service, which require human intervention. AI systems would also offer greater insights that would enable financial institutions to manage their operations optimally in order to make more lucrative decisions.

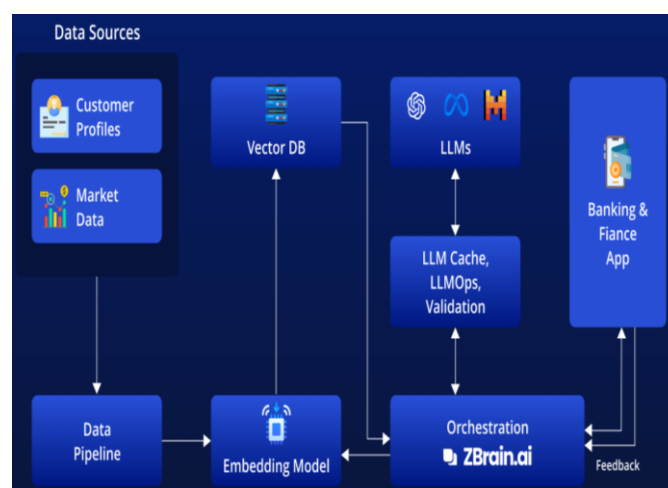


Figure 5: An example of data source integration, vector databases, LLMs, and coordination to achieve effective AI-based solutions in banking and finance applications.

Figure 5 below highlights the design of an AI-based banking and finance application system architecture. The figure indicates how the customer profiles and market data are integrated as the main data sources, which are then moved to a vector database, where they are effectively stored and retrieved. A query is resolved by passing the data through an embedding model to produce vector embeddings to

improve query resolution [38]. These embeddings are analyzed and processed by the LLMs (Large Language Models) in an effort to make a decision. The system is staged by ZBrain.ai, which works with such tasks as LLM cache, LLM operations (LLMops), and validation. The banking & finance app feedback loop enhances the system, making the performance optimized. The architecture significantly benefits the efficiency of operations since it avoids dependence on manual operations and boosts the speed of decision-making in housing finance.

Improved customer experience that can also be obtained through the integration of AI is also a significant advantage. Customers can get quicker responses and more precise information because of AI systems that can reduce the time spent on query handling by up to 40%. This increase in service quality has a direct relationship with increased customer satisfaction. Contented consumers would be more loyal towards the financial establishments and would be more willing to refer others to the services, thus, it could help to increase customer retention rates and general business development. By helping to accelerate the response and increase its accuracy, AI systems also allow financial institutions to deliver personalized services, thus improving the customer experience to an even greater extent [39].

Findings of this experiment make it clear that the introduction of AI-powered systems into housing finance can help immensely to enhance the efficiency of operations, prediction quality, and customer satisfaction. Though challenges are involved in terms of scalability and initial program implementation cost, the benefits of this are long-term and by far outshine the initial obstacles. By adopting AI in their financial services, financial institutions can be in a good position to succeed in a more competitive and data-driven world.

6. Future Research Recommendations

6.1 Scalability

One potential future study is the scalability of AI systems on a global market. The larger the amount of data and the range of variability, especially in organizations that provide housing financial services on a large scale, the bigger the AI system should be able to manage extensive and complicated datasets. The existing models are good in a smaller-scale setting, yet extending such models to different regions with distinct market characteristics and other data is a challenge [40]. Further studies are needed to address the limitation by developing models of AI models that can support and process high-dimensional and multi-source heterogeneous data at the same performance and accuracy as small-scale use.

Banks that work in several countries need to have systems that are able to address local variations in the price of property, the type of customers, and the economic environments. This requires studies on distributed systems that would be convenient in handling the storage and processing of information in various geographical locations. Cloud computing and edge computing systems can also make AI usage more flexible and scalable to ensure that, as the system grows to service the needs of global financial institutions, it can be adjusted accordingly [41]. The future of AI in housing finance is in the domain of adapting to new data patterns in real-time, which enables AI systems to continue remaining effective in various and large-scale data sets.

6.2 AI Integration

The other potential field of research in the future is to combine reinforcement learning and Large Language Models (LLMs) to enhance risk management and adaptive decision-making. Reinforcement learning (RL) allows an AI-based system to optimize decisions over time through learning from interactions with its environment. It is especially applicable in dynamic markets such as housing finance. The addition of reinforcement learning to the LLMs may enable financial institutions to develop AIs that adjust their strategies according to market dynamics, consumer actions, and

changing data trends. For example, AI would be able to keep updating its loan issuance requirements or house pricing algorithms with real-time data, enhancing decision-making precision [42].

Integration of RL and LLMs may also be effective to expand responsiveness, because AI can refresh its models with each new action, subsequent to an action, and its results. This would ensure improved risk management because AI systems will be more resilient to react to unexpected market changes. For example, when an AI system identifies a rising housing market bubble or a major change in customer preferences, the concept of reinforcement learning may allow the system to alter its risk assessment and forecast independently of human interference. The studies should look into how these technologies can be blended together to enhance the decision-making process long-term and to reduce the risk level involved and the overall accuracy of the housing financial services.

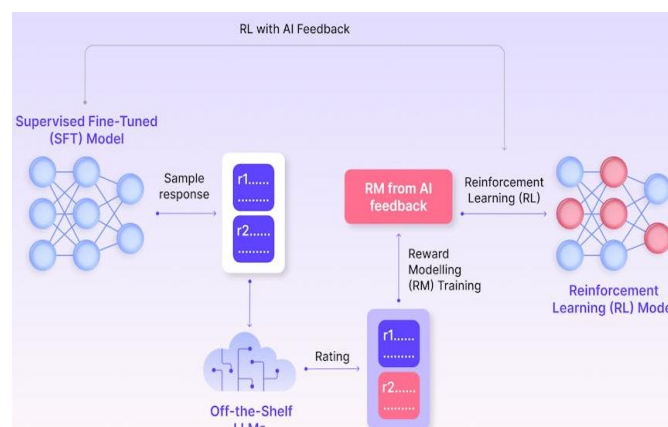


Figure 6: A summary of adopting Reinforcement Learning (RL) and Reforming Large Language Models (LLMs) to adaptive decision-making and risk management in financial services.

Figure 6 below demonstrates how Reinforcement Learning (RL) can be combined with the Large Language Models (LLMs) to enhance decision-making in dynamic environments. This works by beginning with a Supervised Fine-Tuned (SFT) model, which produces sample responses. They are then prompted through the LMS, which takes AI feedback to operate reward models (RM). The RL model takes time to refine during interactions and feedback loops, which makes the system streamline the decisions [43]. This hybridization enables continuous adaption of financial models, including the algorithms to issue loans and the price of properties according to market dynamics. The model has the capacity to modulate risk measurement and predictions on its own and gives better risk management and adaptive decision-making in housing finance, and more so in uncertain markets. Both RL and LLM will make it more resilient and sensitive to market changes and customer behavior.

6.3 Ethical Considerations

With the increasing AI deployed in housing finance, the ethical aspects of bias and the fairness of AI models are paramount. Regulatory frameworks play a major role in the housing finance sector, with an interest in providing equal access to housing and finances [44]. Without a proper design, AI systems can be a source of biases in historical data, discovering and supporting their discrimination based on race, gender, or socioeconomic status. Future studies should proceed to devise methods of identifying and eliminating bias in AI models that are applicable in valuing property, granting loans, and profiling customers. For example, AI-based price predictors of houses would accidentally represent past trends of unjust practices, resulting in incorrect or discriminative valuations. This is particularly problematic in the area of housing finance, where there must be fairness in pricing and granting loans.

To resolve these problems, academics are encouraged to work on bias-conscious algorithms and introduce fairness constraints into AI models. Even the opposing techniques like adversarial debiasing

or fairness-enhancing algorithms can assist in ensuring that the decisions delivered by AI do not reinforce the detrimental societal values. Besides this, AI decision-making needs to be transparent. Future studies should also address ways of increasing accountability and explainability in AI systems, particularly in cases where they will be on high-stakes decisions, such as property values and loan issuance. Making AI systems more interpretable will help the institution guarantee the transparency of decision-making processes and will earn the trust of potential customers and regulatory bodies. Research should also be conducted on the ethical issues of applying AI to credit rating and loan provision models, with a particular emphasis on making sure that the algorithms are not discriminatory, in general, and based on marginalized groups in particular.

Future studies regarding the scaling of AI systems in global markets, incorporating reinforcement learning to enhance adaptive decision-making, and tackling the ethical issues of bias and fairness are areas that need to be considered in light of further AI advancement in the housing finance arena. Scientific studies on distributed computing, cloud computing infrastructure, and algorithms to promote fairness will reduce their optimization so that AI systems can be scaled efficiently and keep their accuracy, efficiency, and fairness optimized in various regions and markets [45]. With the solution to these problems, AI systems of the future will be in a position to provide more dependable, ethical, and scalable solutions to the housing finance sector, ultimately benefiting financial institutions and their clientele.

7. Conclusions

This study has established the potential revolutionary influence of the combination of Large Language Models (LLM) and vector databases in the financial services sector of the housing sector. The study focused on improving two very important aspects, such as speed in query retrieval, as well as the accuracy of property valuation, which are vital towards effective decision-making in housing finance. The findings indicate that query response time costs were reduced by up to 90% with the help of LLMs and vector databases, decreasing by at least 3-5 seconds to 0.5-1 seconds on average. This reduction of time enables financial services institutions to make quicker decisions, which would be essential in time-intensive operations such as loan issuances, market reviews, and house evaluations. Besides enhancing speed, the study has observed that such AI-powered systems also improved the accuracy of predictions. Predictions regarding property prices, or the core of risk assessment and loan provider, were enhanced by 15%, with the AI models reaching 92% accuracy relative to 85% when the conventional approach is used. This improved level of accuracy comes in especially handy when it comes to countermeasures against risks in the housing financial sector, since the enhanced quality of property value assessment lowers the possibility of losses and incidences of poor decision-making. The resultant increase in the speed of data retrieval and accuracy will allow more specific financial forecasting and risk management that are crucial to success in the housing market.

These findings have a significant implication in the real world. To start with, the combination of LLMs and vector databases leads to a 25% drop in operation costs. The reason behind this saving is that manual intervention is minimized in retrieving data and handling the responses of the customers as a result of AI automation. Having AI systems that respond to queries swiftly and correctly, financial organizations can allocate resources more effectively, thus limiting human control over financial operations and operational inefficiencies. These savings also lead to a more competitive business model, particularly in market that requires real-time decision-making. The increased speed of response and the precision of predictions lead to an improvement in customer satisfaction. Enhancing customer experiences by 40% faster responses to customer queries, depending on AI-focused systems, improves customer experience, results in increased client retention rates. Satisfied customers will feel that they can trust their financial institution, refer services to other willing customers and create business

relationships, which are long term. Such a better customer experience has become essential in the housing finance industry, where trust and timely service are critical in ensuring competitive advantage.

The results of the study indicate that LLM and vector databases have considerable value to the housing financial services industry. Such technologies are not only effective in terms of the efficiency of operations, decreasing the queries and increasing the accuracy of the prediction, but also aid in significant cost savings. With the banks still struggling with the problem of large-scale data processing, AI technologies provide a scalable solution to the increasing housing market demands. With the integration of AI, there are improved decision-making processes, customer satisfaction, and financial forecasting that are all important in lasting success in this competitive industry. Future studies must aim at solving the challenges of scalability, ethical issues, and additional optimization of AI systems to enable the support of global markets.

References;

- [1] Brummer, A. (2024). The data needs and solutions of a living real estate company. <https://aaltodoc.aalto.fi/bitstreams/1564743a-6320-4c89-ac60-277853829cce/download>
- [2] Rusum, G. P., & Anasuri, S. (2024). Vector Databases in Modern Applications: Real-Time Search, Recommendations, and Retrieval-Augmented Generation (RAG). *International Journal of AI, BigData, Computational and Management Studies*, 5(4), 124-136.
- [3] Kothandapani, H. P. (2023). Emerging trends and technological advancements in data lakes for the financial sector: An in-depth analysis of data processing, analytics, and infrastructure innovations. *Quarterly Journal of Emerging Technologies and Innovations*, 8(2), 62-75.
- [4] Samala, S. (2024). *Real-time Jira analytics: Integrating JQL with Power BI/Snowflake for predictive agile metrics*. SciPubHouse. <https://scipubhouse.com/home/international-journal-of-sustainability-and-innovation-in-engineering-ijse/content/ijse-2024/real-time-jira-analytics-integrating-jql-with-power-bi-snowflake-for-predictive-agile-metrics/>
- [5] Sukumaran, A. (2023). *Database Design and Modeling with Google Cloud: Learn database design and development to take your data to applications, analytics, and AI*. Packt Publishing Ltd.
- [6] Han, Y., Liu, C., & Wang, P. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. *arXiv preprint arXiv:2310.11703*.
- [7] Karanikolas, N., Manga, E., Samaridi, N., Tousidou, E., & Vassilakopoulos, M. (2023, November). Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics* (pp. 278-290).
- [8] Fatima, S. (2022). The Impact Of Artificial Intelligence On Intellectual Property Laws.
- [9] Pan, J. J., Wang, J., & Li, G. (2024). Survey of vector database management systems. *The VLDB Journal*, 33(5), 1591-1615.
- [10] Hariharan, R. (2024). *API gateway threat prevention in large-scale applications*. SciPubHouse. https://scipubhouse.com/wp-content/uploads/2024/10/011-API_gateway_threat_prevention_in_large-scale_applications.pdf.
- [11] Olayinka, O. H. (2021). Big data integration and real-time analytics for enhancing operational efficiency and market responsiveness. *Int J Sci Res Arch*, 4(1), 280-96.
- [12] Tantalaki, N., Souravlas, S., & Roumeliotis, M. (2020). A review on big data real-time stream processing and its scheduling techniques. *International Journal of Parallel, Emergent and Distributed Systems*, 35(5), 571-601.
- [13] Singh, P. N., Talasila, S., & Banakar, S. V. (2023, December). Analyzing embedding models for embedding vectors in vector databases. In *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)* (pp. 1-7). IEEE.

- [14] Chandra, B., Preethika, P., Challagundla, S., & Gogireddy, Y. End-to-End Neural Embedding Pipeline for Large-Scale PDF Document Retrieval Using Distributed FAISS and Sentence Transformer Models. *Journal ID*, 1004, 1429.
- [15] Zulkifli, A. (2023). Accelerating database efficiency in complex IT infrastructures: Advanced techniques for optimizing performance, scalability, and data management in distributed systems. *International Journal of Information and Cybersecurity*, 7(12), 81-100.
- [16] Faizal, A., & Aisyah, N. Innovative Approaches to Enterprise Database Performance: Leveraging Advanced Optimization Techniques for Scalability, Reliability, and High Efficiency in Large-Scale Systems. *Reliability, and High Efficiency in Large-Scale Systems*.
- [17] Sandhu, A. K. (2021). Big data with cloud computing: Discussions and challenges. *Big Data Mining and Analytics*, 5(1), 32-40.
- [18] Mathur, P. (2024). Cloud computing infrastructure, platforms, and software for scientific research. *High Performance Computing in Biomimetics: Modeling, Architecture and Applications*, 89-127.
- [19] Ooi, B. C., Cai, S., Chen, G., Shen, Y., Tan, K. L., Wu, Y. & Zhao, Z. (2024). NeurDB: an AI-powered autonomous data system. *Science China Information Sciences*, 67(10), 200901.
- [20] Nwachukwu, G. (2024). Enhancing credit risk management through revalidation and accuracy in financial data: The impact of credit history assessment on procedural financing. *International Journal of Research Publication and Reviews*, 5(11), 631-644.
- [21] Ionescu, S. A., & Diaconita, V. (2023). Transforming financial decision-making: the interplay of AI, cloud computing and advanced data management technologies. *International Journal of Computers Communications & Control*, 18(6).
- [22] Ismanov, I., Qayumov, N., Mukhamadjonova, D., & Akhmadaliyev, B. (2024, April). AI and cost management: Strategies for reducing expenses and improving profit margins in business. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (Vol. 1, pp. 1-7). IEEE.
- [23] Gomes, F. C. (2024). Generating process descriptions from BPMN models: towards an approach with Large Language Models.
- [24] Gadde, H. (2023). Leveraging AI for Scalable Query Processing in Big Data Environments. *International Journal of Advanced Engineering Technologies and Innovations*, 1(02), 435-465.
- [25] Ojika, F. U., Owobu, W. O., Abieba, O. A., Esan, O. J., Ubamadu, B. C., & Daraojimba, A. I. (2022). The Role of Artificial Intelligence in Business Process Automation: A Model for Reducing Operational Costs and Enhancing Efficiency.
- [26] Krauze, A. V., Zhuge, Y., Zhao, R., Tasci, E., & Camphausen, K. (2022). AI-driven image analysis in central nervous system tumors-traditional machine learning, deep learning and hybrid models. *Journal of biotechnology and biomedicine*, 5(1), 1.
- [27] Nagaraj, V. (2024). *Addressing power efficiency challenges in AI hardware through verification*. SciPubHouse.
<https://scipubhouse.com/home/international-journal-of-sustainability-and-innovation-in-engineering-ijsie/content/ijsie-2024/addressing-power-efficiency-challenges-in-ai-hardware-through-verification/>
- [28] Oetama, S., Susanto, H., & Rizwannur, W. (2024). Effect Of Online Tracking System And Delivery Timeliness On Customer Satisfaction (Case Study On J & T Express Sampit). *International Journal of Science, Technology & Management*, 5(4), 962-969.
- [29] Singh, P., & Singh, V. (2024). The power of AI: enhancing customer loyalty through satisfaction and efficiency. *Cogent Business & Management*, 11(1), 2326107.
- [30] Hsu, C. L., & Lin, J. C. C. (2023). Understanding the user satisfaction and loyalty of customer service chatbots. *Journal of Retailing and Consumer Services*, 71, 103211.

- [31] Cascio, N. (2024). Rental Riches: Unlocking Wealth Through Mid-Term and Corporate Properties.
- [32] Sarker, I. H. (2022). AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN computer science*, 3(2), 158.
- [33] Hammad, A., & Abu-Zaid, R. (2024). Applications of AI in decentralized computing systems: harnessing artificial intelligence for enhanced scalability, efficiency, and autonomous decision-making in distributed architectures
- [34] Panwar, V. (2024). AI-driven query optimization: Revolutionizing database performance and efficiency. *Int. J. Comput. Trends Technol*, 72, 18-26.
- [35] Ahmad, Z., Rahim, S., Zubair, M., & Abdul-Ghafar, J. (2021). Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review. *Diagnostic pathology*, 16(1), 24.
- [36] Avickson, E. K., Omojola, J. S., & Bakare, I. A. (2024). The role of revalidation in credit risk management: ensuring accuracy in borrowers' financial data. *Int J Res Publ Rev*, 5(10), 2011-2024.
- [37] Bhatnagar, S., & Mahant, R. (2024). Unleashing the Power of AI in Financial Services: Opportunities, Challenges, and Implications. *Artificial Intelligence (AI)*, 4(1).
- [38] Naseri, S., Dalton, J., Yates, A., & Allan, J. (2021, March). Ceqe: Contextualized embeddings for query expansion. In *European conference on information retrieval* (pp. 467-482). Cham: Springer International Publishing.
- [39] Rahmani, F. M., & Zohuri, B. (2023). The transformative impact of AI on financial institutions, with a focus on banking. *Journal of Engineering and Applied Sciences Technology. SRC/JEAST-279*. DOI: [doi.org/10.47363/JEAST/2023\(5\),192,2-6](https://doi.org/10.47363/JEAST/2023(5),192,2-6).
- [40] Bjarghov, S., Löschenbrand, M., Saif, A. I., Pedrero, R. A., Pfeiffer, C., Khadem, S. K., ... & Farahmand, H. (2021). Developments and challenges in local electricity markets: A comprehensive review. *IEEE Access*, 9, 58910-58943.
- [41] Haefner, N., Parida, V., Gassmann, O., & Wincent, J. (2023). Implementing and scaling artificial intelligence: A review, framework, and research agenda. *Technological Forecasting and Social Change*, 197, 122878.
- [42] Xiang, X., Xue, J., Zhao, L., Lei, Y., Yue, C., & Lu, K. (2024, June). Real-time integration of fine-tuned large language model for improved decision-making in reinforcement learning. In *2024 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
- [43] Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (2020). Optimizing Decision-Making with AI-Enhanced Support Systems: Leveraging Reinforcement Learning and Bayesian Networks. *International Journal of AI and ML*, 1(2).
- [44] AlQahtany, A. M. (2022). Government regulation and financial support on housing delivery: lessons learned from the Saudi experience. *International Journal of Housing Markets and Analysis*, 15(3), 613-631.
- [45] Ghelani, D. (2024). Optimizing resource allocation: Artificial intelligence techniques for dynamic task scheduling in cloud computing environments. *International Journal of Advanced Engineering Technologies and Innovations*, 3(1), 132-156.