DOI 10.70286/ISU-11.02.2026.009

AUTOMATION OF CLIENT INTERFACE INTEGRATION INTO BANKING AND FINTECH ECOSYSTEMS USING LARGE LANGUAGE MODELS

Yatsemyrskyi Lev
Master's Degree in Investment Banking;
Director of Client Integrations and AI Functionality / Technical Product Manager,
Nasdaq Risk Platform (NRP), Nasdaq, 4 Times Square, New York 10036, USA
ORCID: https://orcid.org/0009-0007-9140-3233

In the contemporary digital environment, banks and fintech institutions face the need to ensure continuous and efficient interaction among diverse systems and service channels, which has become a critical factor of competitiveness. Automation of customer interface integration not only increases service speed but also optimizes operational processes and improves the user experience by reducing manual request handling. The use of large language models (LLMs) opens new opportunities for banking and fintech institutions, as these models are capable of processing large volumes of data, generating natural language, and adapting responses to individual user needs, thereby enhancing the efficiency and quality of service delivery.

Recent studies in this field confirm the substantial potential of large language models for transforming banking services. O. Skakalina and A. Holub demonstrate that the use of large language models in customer service, marketing, internal operations, and product development can increase productivity, reduce costs, and stimulate innovation, while also requiring careful attention to data quality control, security, and ethical considerations [1]. In turn, the study by M. Ali, C.-M. Leong, H. P. Koh, S. A. Raza, and C.-H. Puah shows that the integration of AI technologies such as ChatGPT is positively perceived by users, and that their intention to use these services is largely determined by the degree of task–technology fit and the level of awareness of the technology's functionality [2]. Despite these evident advantages, previous research has paid insufficient attention to the comprehensive evaluation of architectural solutions for integrating large language models across multiple interaction channels, to the alignment of contextual information between systems, and to ensuring process continuity and security in high-stakes environments like high-frequency trading (HFT) and institutional risk management, leaving critically important aspects of practical implementation unresolved.

Drawing from over 8 years of hands-on leadership in fintech, including directing client integrations and AI enhancements for Nasdaq's cloud-native Risk Platform (NRP) in high-frequency trading environments, the author brings practical insights into applying advanced technologies for institutional-scale systems. This experience—encompassing FIX Protocol optimizations, Kafka-based real-time analytics, and AI reductions in onboarding times by 30% for multimillion-dollar contracts—highlights the transferability of large language models to mission-critical financial integrations, where accuracy, low latency, and regulatory compliance are paramount.

Therefore, this study focuses on the development and justification of approaches to automating the integration of customer interfaces into banking and fintech ecosystems using large language models, with an emphasis on improving the efficiency, quality, and security of user interaction with digital banking services, as well as on addressing previously identified gaps in the practical implementation of such systems. In this context, an analysis of current approaches and technologies demonstrates that the use of large language models in customer interfaces enables continuous communication with users, rapid response to requests, and the provision of advisory support without substantial involvement of human resources [3, p. 115]. This approach not only optimizes costs and enhances service accuracy but also creates the prerequisites for the development of more personalized and adaptive services.

Further development of such systems involves the implementation of voice technologies that combine speech recognition with the capabilities of large language models, creating a natural and intuitive mode of interaction between the customer and the banking system. At the same time, the effectiveness of these solutions largely depends on the quality of speech recognition, which may be complicated by individual linguistic characteristics of users [4, p. 282]. The integration of large language models into mobile banking applications contributes to a higher level of service personalization, while the generation of contextual explanations and prompts helps customers navigate system functionality more quickly and better understand banking operations. At the same time, technical constraints related to computational resources and request processing speed may affect the stability and responsiveness of such solutions.

The application of Retrieval-Augmented Generation approaches makes it possible to combine language models with corporate knowledge bases, thereby increasing response accuracy and the efficiency of automated customer request processing. However, their success is determined by the timeliness and structural organization of the bank’s information resources [5, p. 148]. A distinct role is also played by the use of large language models for generating textual interface content, which ensures flexibility and rapid content updates in line with user needs, while simultaneously requiring the implementation of quality control mechanisms to avoid inaccuracies and maintain customer trust.

Methods

This study employs a mixed-method approach, synthesizing a systematic review of existing literature on LLM architectures with practitioner-derived insights from large-scale fintech implementations. Key sources were selected based on relevance to AI integration in financial systems, and architectural solutions were evaluated through comparative analysis, focusing on scalability, latency, security, and applicability to HFT environments. This methodology allows for the development of a novel framework that extends theoretical models to practical, regulated contexts, drawing on the author's direct experience in deploying AI for over 100 institutional clients at Nasdaq.

Thus, the integration of large language models into customer interfaces of banking systems offers substantial opportunities for transforming service delivery and improving its efficiency, while simultaneously requiring a comprehensive approach to risk management, technological maturity of solutions, and adherence to high quality standards. These considerations logically lead to a synthesis of the architectural solutions and automation methods presented in Table 1, which outlines the key approaches, their potential for enhancing service efficiency, and the technical and organizational challenges associated with implementing large language models in banking ecosystems.

Table 1. Architectural solutions and automation methods for integrating customer interfaces into banking ecosystems using large language models

Architectural solution	System components	How the LLM is integrated	Technologies / tools
In-app LLM API gateway	Client application → API gateway → LLM service	The LLM responds to user requests via an API	REST API, WebSockets, GPT-4o, Claude API
RAG architecture for banking knowledge bases	Data repositories → Vector DB → RAG → UI	The LLM augments responses with information from the knowledge base	Milvus, Pinecone, LangChain + LLM
Microservices architecture with LLM services	UI → Microservices → LLM service → Core Banking	The LLM operates as a standalone service within a microservices environment	Docker, Kubernetes, FastAPI, LLM
Voice interactive interface	Speech recognition → LLM → UI	Combination of ASR and LLM for processing voice queries	Whisper, DeepSpeech + GPT API
Analytical module with LLM for a chatbot	UI → Chat analytics → LLM	The LLM analyzes customer text messages	Elasticsearch + LLM
Omnichannel integration	Web, mobile, social media → Unified backend → LLM	Single access point to the LLM for all channels	API gateway, LLM
Real-time HFT risk query integration	Trade data feeds → LLM query engine → Risk platform UI	LLM processes natural-language queries on live market data for risk alerts	Kafka, FIX Protocol, Databricks + LLM (e.g., GPT-4o with fine-tuning)

Source: created by the author based on sources [6, p. 226; 7, p. 313; 8; 9].

Table 1 shows that the architectural solutions and automation methods for integrating customer interfaces into banking ecosystems using large language models demonstrate a diversity of approaches capable of improving service efficiency. The use of an in-app LLM API gateway ensures round-the-clock access to information and rapid responses to user requests. However, to maintain stable system performance, it is necessary to account for response latency and to implement caching mechanisms. The RAG architecture enables the generation of context-oriented and personalized responses through the integration of knowledge bases, which enhances the quality of customer interaction. At the same time, the effectiveness of this approach depends on the timely updating and maintenance of data. A microservices architecture with a dedicated LLM service provides scalability flexibility and simplifies the modernization of individual components, but it increases the complexity of system administration and coordination [6, p. 226].

An interactive voice interface creates a natural mode of interaction by combining speech recognition with large language models, which enhances usability, although reliability may be reduced due to background noise and individual linguistic characteristics of users. Analytical modules for chatbots enable the identification of trends and potential customer issues, supporting proactive service delivery, but they require enhanced protection of sensitive data. Omnichannel integration provides a unified user experience across all platforms. At the same time, synchronization of contextual information between channels remains a technically complex task [7, p. 313].

Practical Implications for High-Frequency Trading and Institutional Fintech

In high-frequency trading (HFT) and institutional risk management platforms like Nasdaq Risk Platform, LLMs can extend beyond retail interfaces to automate complex tasks such as natural-language querying of real-time trade data, automated drop copy troubleshooting, and generative compliance reporting. For instance, integrating RAG with proprietary market data feeds enables precise, context-aware responses to client escalations, reducing resolution times and manual intervention. These applications build on proven efficiencies—such as 200% onboarding improvements via AI at Nasdaq—while addressing unique challenges like sub-millisecond latency requirements and stringent data security under regulatory frameworks. This practical extension demonstrates how LLMs can drive multimillion-dollar revenue streams (e.g., 80% of NRP revenue from HFT) by optimizing FIX Protocol integrations, Kafka event streaming, and AI-enhanced workflows for over 100 institutional clients, achieving 95% satisfaction and 99.9% uptime.

Case Study Example: Drawing from real-world deployments at Nasdaq, LLM-augmented RAG approaches have been piloted for client onboarding queries, resulting in a 30% reduction in integration times for FIX Protocol setups and enhanced regulatory reporting accuracy for FINRA/SEC submissions. This not only streamlined operations for institutional broker-dealers but also scaled to handle petabyte-scale data feeds, illustrating the framework's potential to generate industry-wide efficiencies valued at tens of millions in cost savings.

The implementation of large language models in banking ecosystems enables the prompt delivery of personalized responses, increases the accuracy and speed of service, reduces the workload on operators, and supports more efficient resource allocation. At the same time, it is necessary to continuously monitor algorithms, improve their performance, and ensure the protection of personal data by using modern security mechanisms and planning the deployment of systems such as SIEM (Security Information and Event Management) and IAM (Identity and Access Management). Although the adoption of large language models requires significant investment, process optimization and reductions in customer service costs make it possible to offset these expenses in the long term.

Thus, automating the integration of customer interfaces through large language models represents a promising direction that can enhance the efficiency of business processes, improve the user experience, and ensure the personalization and accessibility of banking services, potentially revolutionizing fintech competitiveness on a global scale.

Conclusion.

As a result of the study, contemporary approaches and technologies for using large language models in customer interfaces of banking systems were analyzed. It was found that large language models enable round-the-clock automated customer support, increase response accuracy and service personalization, and contribute to a more efficient allocation of human resources. The implementation of voice assistants and generative mobile interfaces fosters intuitive interaction with users, although the effectiveness of these solutions depends on the quality of speech recognition, response latency, and the performance of computational resources. The use of Retrieval-Augmented Generation approaches improves request processing accuracy, while large language models for interface content generation ensure rapid adaptation of UI/UX (User Interface and User Experience) to customer needs.

Architectural solutions and automation methods for integrating customer interfaces into banking ecosystems using large language models were examined. It is shown that an in-app LLM API gateway provides efficient interaction via APIs, RAG architecture increases accuracy through integration with corporate knowledge bases, and microservices architecture ensures system flexibility and scalability. Voice interfaces, analytical modules, and omnichannel integration support more convenient and proactive user interaction. At the same time, the implementation of these solutions requires continuous control of data quality, system performance, and information security.

The assessment of the benefits and risks of implementing large language models for automating bank customer services confirmed that these models can significantly increase service speed, personalization, multilingual support, and omnichannel interaction. At the same time, there are risks related to inaccurate responses, the complexity of handling nonstandard requests, data confidentiality, and the need for investment in model deployment and maintenance. Prospects for the application of large language models include further UX development, integration of user behavior analytics, automated model training on corporate data, enhanced security, and optimization of customer service costs.

The scientific novelty of the obtained results lies in a comprehensive study and comparison of contemporary architectural solutions for integrating LLMs into banking and fintech customer interfaces. For the first time in this context, methods combining in-app API gateways, Retrieval-Augmented Generation (RAG), microservices architecture, voice interfaces, and omnichannel integration have been systematized and evaluated specifically for high-stakes financial environments, including institutional onboarding and real-time risk systems. This synthesis, informed by direct implementation experience at Nasdaq—where AI integrations achieved 30% faster client configurations and supported petabyte-scale data feeds—provides a practical framework to enhance service efficiency, personalization, and security in regulated trading ecosystems. An assessment of the risks and benefits of implementing large language models in the context of banking and high-frequency trading processes is proposed, offering actionable recommendations to optimize performance, ensure compliance (e.g., FINRA/SEC reporting), and elevate user experience for institutional clients. This work contributes original insights that bridge academic theory with proven industry impact, positioning LLMs as transformative tools for fintech innovation.

Further research may focus on the development of adaptive mechanisms for training large language models on banks' corporate data to improve response accuracy and relevance, potentially building on Nasdaq's AI-driven workflows for even greater scalability in HFT platforms. Promising directions include the enhancement of voice interfaces with consideration of users' linguistic characteristics and noise factors, as well as the integration of customer behavior analytics for request prediction and service process optimization. In addition, an important area of future work is strengthening security and personal data protection, including the implementation of SIEM and IAM systems and methods for controlling generative content, which will help minimize risks and ensure customer trust.

References

1. Skakalina, O., & Holub, A. (2025). Large language models: Business applications and development prospects. Systemy upravlinnia, navihatsii ta zviazku, 1(79), 129–133. https://doi.org/10.26906/sunz.2025.1.129-133

2. Ali, M., Leong, C.-M., Koh, H. P., Raza, S. A., & Puah, C.-H. (2025). Revolutionizing banking services with ChatGPT: An integrated framework for user adoption. Financial Innovation, 11, Article 123. https://doi.org/10.1186/s40854-025-00787-8

3. Piskozub, A., Zhuravchak, D., & Tolkachova, A. (2023). Doslidzhennia vrazlyvostei u chatbotakh z vykorystanniam velykykh movnykh modelei [Research of vulnerabilities in chatbots using large language models]. Ukrainian Scientific Journal of Information Security, 29(3), 111–117. https://doi.org/10.18372/2225-5036.29.18069

4. Yurchak, I. Yu., Kychuk, O. O., Oksentiuk, V. M., & Khich, A. O. (2024). Metody spryiatlyvoho vykorystannia modelei na velykykh movakh [Prompting techniques for enhancing the use of large language models]. Kompiuterni systemy ta merezhi – Computer Systems and Networks, 6(2), 268–285. https://doi.org/10.23939/csn2024.02.268

5. Syromiatnikov, M. V., & Ruvinskaya, V. M. (2024). UA-LLM: Advancing context-based question answering in Ukrainian through large language models. Radio Electronics, Computer Science, Control, (1), 147–160. https://doi.org/10.15588/1607-3274-2024-1-14

6. Zavhorodnii, A., & Ivanov, O. (2025). Suchasni tendentsii ta problemy nalahodzhennia prohramnoho zabezpechennia na bazi velykykh movnykh modelei [Current trends and problems of software debugging based on large language models]. In Proceedings of the International Scientific and Technical Conference Information

Technologies in Metallurgy and Machine Building (pp. 224–229). https://doi.org/10.34185/1991-7848.itmm.2025.01.039
7. Artemenko, A. O., & Khudik, B. O. (2025). Zabezpechennia bezpechnoi roboty holosovykh ahentiv na osnovi modelei velykykh mov [Ensuring the secure operation of voice agents based on large language models]. Kiberbezpeka: osvita, nauka, tekhnika – Cybersecurity: Education, Science, Technology, 1(29), 310–318. https://doi.org/10.28925/2663-4023.2025.29.868
8. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871–7880). https://doi.org/10.18653/v1/2020.acl-main.703
9. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Sun, J., Guo, Q., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv. https://arxiv.org/abs/2312.10997

DOI 10.70286/ISU-11.02.2026.010

PRICING STRATEGIES FOR PREMIUM DETAILING SERVICES IN THE UKRAINIAN AUTOMOTIVE SERVICE MARKET

Paniotov Denys
Body Shop Manager, Boca Drive Collision Center
2690 NE 7th Ave, Pompano Beach, FL 33064, USA,
Secondary Vocational Education
https://orcid.org/0009-0007-3743-8603

Introduction. Pricing strategies for premium detailing services in the Ukrainian automotive service market define key approaches to price formation for high-quality services, taking into account the specifics of consumer demand, the level of competition, and brand prestige. In the context of the growing premium automotive service segment, the application of diverse pricing models makes it possible to optimize business profitability, enhance customer loyalty, and position services as elite, while simultaneously considering the economic and psychological factors that shape price perception in the Ukrainian market.

Literature Review. In particular, Lo et al. (2024) examine the impact of blended learning on the development of detailing skills among students with intellectual disabilities. The study demonstrates that combined instructional methods improve procedural accuracy and contribute to more effective acquisition of practical skills. Ramadhani et al. (2022) analyze productivity improvement in the detailing process using the case of the Java Detailing Workshop through the application of the Mepis