UDC 502.55:504.5:576.8:631.46
DOI https://doi.org/10.32689/maup.it.2025.2.32

Sofiia SHVETS
Master of Science, Junior Research Scientist (Computer Systems),
Applied Programmer, Data Scientist, NinjaTech AI
***ORCID:** 0009-0000-7896-6479*

INTEGRATION OF LARGE LANGUAGE MODELS INTO CHATBOT-BASED CUSTOMER CALL PROCESSING SYSTEMS

Abstract. Purpose of the work. *To propose a hybrid model of automated text customer service that combines semantic classification with the generative capabilities of LLM to improve the accuracy, relevance, and naturalness of responses.*

Methodology. *A system has been developed that selects one of three response mechanisms depending on the classification of the intent and emotional tone of the query: template rules, search matching, or LLM generation. Experimental verification was performed on a corpus of 7,500 queries (authentic and synthetic); evaluation was conducted using BLEU, ROUGE-L, and expert criteria for comprehensibility, naturalness, and user trust.*

Scientific novelty. *For the first time, it has been shown that semantic routing in conjunction with LLM forms a more robust and adaptive system capable of correctly processing complex or emotionally charged queries. The proposed model outperformed baseline approaches in terms of accuracy (92.1%), BLEU 0.78, ROUGE-L 0.81, and the lowest failure rate, and also received the highest expert ratings.*

Conclusions. *The hybrid model reduces operator workload, increases user satisfaction, and easily adapts to customer behavior dynamics, providing empathetic and effective responses. Its practical value has been confirmed by examples from e-commerce, banking, healthcare, and public services. Implementation challenges include integration with legacy systems, regular knowledge base updates, and moderation of generated content. Further research is focused on personalization, multimodal interaction, active learning, and optimization of computing resources, laying the foundation for the development of advanced chatbots in areas with a critical need for high-quality automated support.*

Key words: *customer service processing, hybrid model, semantic classification, large language models, BLEU Score, ROUGE-L Score, chatbot, customer service optimization.*

Софія ШВЕЦЬ. ІНТЕГРАЦІЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ У СИСТЕМИ ОБРОБКИ КЛІЄНТСЬКИХ ЗАПИТІВ НА ОСНОВІ ЧАТ-БОТІВ

Анотація. Мета роботи. *Запропонувати гібридну модель автоматизованого текстового обслуговування клієнтів, що поєднує семантичну класифікацію з генеративними можливостями LLM для підвищення точності, релевантності та природності відповідей.*

Методологія. *Розроблено систему, що залежно від класифікації інтенції та емоційного тону запиту вибирає один із трьох механізмів відповіді: шаблонні правила, пошукове зіставлення або генерацію LLM. Експериментальна перевірка виконана на корпусі 7 500 запитів (автентичних і синтетичних); оцінка проведена за BLEU, ROUGE-L та експертними критеріями зрозумілості, природності й довіри користувачів.*

Наукова новизна. *Вперше показано, що семантична маршрутизація у зв'язці з LLM формує більш стійку й адаптивну систему, здатну коректно обробляти складні або емоційно тоновані звернення. Запропонована модель перевищила базові підходи за точністю (92,1 %), BLEU 0,78, ROUGE-L 0,81 та найменшою частотою збоїв, а також отримала найвищі експертні оцінки.*

Висновки. *Гібридна модель зменшує навантаження операторів, підвищує задоволеність користувачів і легко адаптується до динаміки поведінки клієнтів, пропонуючи одночасно емпатичні та дієві відповіді. Практичну цінність підтверджено на прикладах з електронної комерції, банківської сфери, охорони здоров'я та публічних сервісів. Виклики впровадження охоплюють інтеграцію з легасі-системами, регулярне оновлення баз знань і модерацію згенерованого контенту. Подальші дослідження спрямовано на персоналізацію, мультимодальну взаємодію, активне навчання та оптимізацію обчислювальних ресурсів, що закладає основу для розвитку передових чат-ботів у сферах з критичною потребою у високоякісній автоматизованій підтримці.*

Ключові слова: *обробка клієнтських запитів, гібридна модель, семантична класифікація, великі мовні моделі, BLEU, ROUGE-L, чат-бот, оптимізація обслуговування клієнтів.*

Problem statement. Most customers face the chatbots' inability in support services to understand atypical or emotional requests, which increases the number of unresolved requests and worsens the user experience. Traditional rule-based or scripted solutions are losing relevance because of limited flexibility, the inability to adequately respond to atypical or emotionally coloured requests, as well as the difficulty in scaling multilingual support. These limitations reduce customer satisfaction and create additional workload for agents in cases where automated systems cannot cope.

LLMs, such as GPT-4, show potential in solving these problems because of their ability to generate natural and contextually-based responses. However, integrating LLMs into support systems has a number of challenges. These include choosing the right response mode (LLM, rule-based, retrieval) and building an architecture that ensures speed, accuracy, and secure generation. Furthermore, most current studies focus either on the isolated use of LLM or on classic chatbots without a hybrid approach. This creates space for the development of flexible, adaptive architectures that combine the strengths of different approaches.

Review of recent studies and publications. Recently, a significant amount of research has been devoted to the integration of large language models (LLMs) into customer service systems. These models show the potential to improve the quality of dialogues, personalization and efficiency of query processing. For example, Xu et al. in their study [1] presented a new customer service method that combines search-augmented generation with a knowledge graph. This method was applied in the LinkedIn customer service team for six months and reduced the average time to resolve the problem by 28.6%.

Jo & Seo [3] analysed the role of LLMs in modelling the emotional tone of the response, focusing on emotional aspects, but not covering technical or informational requests. Wulf & Meierhofer [11] studied automatic text correction, customer query generalization, and question answering using LLMs. Rüdel & Leidner [10] considered the integration of rule-based systems with neural chatbots, emphasizing the importance of controlling and avoiding "hallucinations" of the models. Kruk et al. [5] presented BanglAssist, a multilingual customer service chatbot capable of handling code-switching and dialect variations, which emphasizes the need for systems to adapt to linguistic diversity.

Hong et al. [2] proposed the Similar Question Generation (SQG) approach based on LLMs. It enables creating a significant number of diverse questions while maintaining semantic consistency with the original question-answer (QA) pair. This is achieved by using the natural language understanding (NLU) capabilities of the master through fine-tuning using specially designed prompts.

Practical implementations also demonstrate the effectiveness of LLM in support services. For example, Klarna implemented AI agents that perform the work of 700 support staff, which indicates significant automation potential. Shopify also used LLM to process basic queries, reducing the workload on operators and improving the speed of responses. In 2024, Duolingo integrated GPT-4 into its platform to improve the user experience. In particular, the Role Play and Explain My Answer functions enable users to have dialogues with AI characters and receive detailed explanations for their answers. This brought the platform closer to the level of a personal tutor. As Bicknell, the project manager, notes: "We've really come close to our ideal of being a personal tutor for every user" [7].

Despite these achievements, there are gaps in research, including:

– Lack of adaptive query processing strategies that take into account the type, complexity, and emotional context of the request.

– Limited attention to multilingualism and dialect variation processing.

– Insufficient implementation of hybrid architectures that combine rule-based, retrieval, and LLM approaches with dynamic switching between them.

This creates space for the development of a new flexible, adaptive chatbot architecture that combines the strengths of different approaches, provides multilingual support, and takes into account the emotional context of requests. This is what this study deals with.

Problem statement. The aim of this study is to develop and experimentally test a hybrid chatbot architecture. It integrates LLMs into the client request processing system, taking into account the type, complexity, and emotional colouring of requests.

Hybrid chatbot architecture. LLMs [8] are a modern advancement in Natural Language Processing (NLP). They are an important component in creating intelligent chatbots, virtual assistance systems, and automated customer support services.

A language model is a statistical or neural system that studies probabilistic patterns of sequences of words or tokens in text. It predicts the next word or generates a text response based on these patterns. The main architectural basis of most modern LLMs are Transformers [9], a type of neural network. They enable efficient processing of large amounts of text data thanks to the Self-Attention mechanism, which ensures that dependencies between all words in the input sequence are taken into account, regardless of their position.

LLMs are trained by supervised learning or self-supervised learning. In the case of self-supervised learning, the model is tasked with predicting hidden or next elements in text data without explicitly labelling examples. The process involves processing large text corpora that can contain milliards of tokens. Intermediate steps teach the model to recognize language structure, logical connections between sentences, contextual dependencies, and even elements of general erudition.

Typical tasks for pre-training are word masking tasks (Masked Language Modelling, MLM) or text autogeneration (Causal Language Modelling, CLM). Despite their results in natural language generation, LLMs have the following limitations [10–11]:

– hallucinations, i.e. fictional facts or generation of incorrect information;

– difficulties in domain adaptation – LLMs may not cope well with specific topics without additional training;

– high computational costs both at the training stage and during model deployment.

The purpose of integrating chatbots with integrated LLMs into customer service systems is to provide automated support with minimal human intervention. However, the specifics of real user requests – including short phrases, unstructured queries, a large number of errors or slang – pose serious challenges to the models. Besides, in a business environment, it is important to ensure the controllability of responses, maintain service quality standards and ensure that responses comply with corporate policies.

The application of LLMs for chatbots covers a variety of industries, including e-commerce, healthcare, banking, education, government services, and technical support. At the same time, the integration of LLMs into such systems requires solving the problems of routing requests, filtering responses, a built-in fact validation mechanism, and optimizing the generation speed.

The creation of hybrid architectures is emphasized among the possible ways to improve the quality of LLM-based chatbots. They combine the capabilities of generative models with pre-processing of incoming requests, semantic classification and the use of specialized rules or knowledge bases. Such solutions increase the accuracy, relevance, and controllability of automated responses.

This study proposes a hybrid model of processing customer requests, which combines several approaches to analysing and generating responses in chatbots. The model is based on the integration of NLP methods, machine learning (ML), and current LLMs. This approach enables its adaptation to a wide range of requests – from typical to non-standard or complex. The model's function consists of several main stages (Figure 1):

1. Receiving a user request (Input): The user sends a text message in a natural language. The system accepts this message as input for further analysis.

2. Request classification (Intent Recognition): at this stage, the system classifies the user's intention. The NLU component – a software module that analyses the grammatical structure of the sentence, keywords and context – is used for this purpose.

3. Response strategy selection: After classifying the request, the system selects the optimal response strategy. This hybrid model provides three main mechanisms: – template responses (Rule-Based);

– retrieval from the knowledge base (Retrieval-Based);

– LLM-based generation (Generative LLM-Based).

4. Formation of the final response depending on the selected strategy (Response Generation).

5. Sending the generated response to the user in the chat (Output).

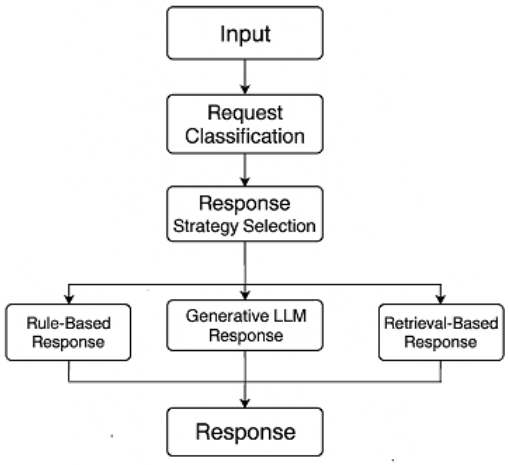

Fig. 1. Model operation diagram

Source: created by the author

This model has several advantages. The combination of three approaches enables its adaptation to different request types. The model can also serve thousands of users simultaneously, changing only the generative component as needed. LLM is involved only when other approaches do not provide sufficient quality of the

response, which helps to save resources. Besides, LLM creates more natural, polite, and contextually relevant responses.

The developed model was compared with the following models:

1. Rule-Based Model [12] This model operates on the basis of hard-coded rules and templates. It relies on predefined keywords, phrases and logical conditions, based on which the chatbot selects the answer.

Advantages: speed, ease of implementation, stability of the response.

Disadvantages: low flexibility, inability to handle complex or new requests, the need to manually update the rules. Usually used in small projects or at the initial stages of customer support automation.

2. Retrieval-Based Model [13] The model is based on the principle of finding the most relevant answer among the previously prepared ones. The user's request is converted into a vector representation (for example, via TF-IDF, FastText or Sentence-BERT), after which it is compared with the vectorized answers in the knowledge base. The most similar answer is returned to the user.

Advantages: good quality for repeated requests, maintaining control over the answers.

Disadvantages: lack of flexibility in formulating queries, poor adaptation to new situations.

This approach is widely used in FAQ systems, banking chatbots, technical support.

3. Generative LLM-based model (LLM-Only) In this model, LLM only is used to generate the answer, such as GPT-3.5, GPT-4 or similar. All logic, classification and formulation of the answer are delegated to the model, which, based on training on huge arrays of text, generates answers that are stylistically and emotionally close to human communication.

Advantages: high flexibility, naturalness of language, ability to solve non-standard situations.

Disadvantages: high consumption of computing resources, slowness in some cases, possible generation of inaccurate or unwanted answers.

Increasingly used in corporate solutions but requires control over the quality of the answer and validation of the content.

The proposed hybrid model combines the strengths of all three approaches: template response efficiency, extraction accuracy, and generation flexibility. Comparison with these models allows for a comprehensive assessment of response quality, performance, stability, and resource consumption.

Course of the experiment. The experiment was conducted on the basis of an internal simulation platform that models the support service of a hypothetical telecommunications provider. The database of requests was formed by combining real user requests – an anonymous history of support service requests (6,000 dialogues). This was supplemented by artificially generated requests created using generative models to cover rare situations, as well as a set of test scenarios developed manually by customer service experts (150 dialogues covering typical cases – password loss, service connection, complaints, billing issues, etc.). In total, the experiment involved a database of 7,500 requests covering 35 main request types.

Some of the requests were simulated through real users for obtaining a qualitative assessment. In particular, the study involved 50 volunteers, who played the role of users with predefined scenarios. This included unexpected wording of requests, language errors, sarcasm, etc. Ten support agents acted as a reference response, answering the same queries without the chatbot. The experiment lasted for two weeks. During this time, each model (Rule-Based, Retrieval-Based, LLM-Only, hybrid) answered the same set of queries.

Generative models were run on a cloud infrastructure with GPU (NVIDIA A100), template and extraction models on CPU servers, chatbots were integrated via API into a simulated support platform. The logging system provided a full record of dialogues and reactions. All requests were fixed in advance (i.e., not dynamically generated). Each model had the same time limit (3 seconds) for a response. LLM responses were additionally filtered for failures.

Methods for evaluating results. The evaluation was carried out using a combination of metrics and expert evaluation. Automatic metrics:

Accuracy – the accuracy of hitting the correct request template. Shows how well the system finds the correct answer immediately (without clarification).

$$Accuracy = \frac{N_r}{N} 100\%,$$

where Nr – the number of correct first answers, N – the total number of requests.

BLEU score – determines the coincidence of text fragments of a certain length (n-grams) between the generated and reference answer.

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n \log(p_n),$$

where BP – penalty for short answer:

$$BP = \begin{cases} 1, & if\ l \geq l_e \\ e^{1-\frac{l_e}{l}} & else \end{cases},$$

where *le* – length of the reference answer, *l* – length of the answer.

pn – accuracy of text fragments, *wn* = 0,25 – weight of each text fragment.

ROUGE-L Score – determines the length of the longest common substring between the answer and the reference answer:

l –pn –wn = 0.25 –ROUGE-L Score –

$$ROUGE - L = \frac{(1+\beta^2) \cdot P \cdot R}{R + \beta^2 \cdot P},$$

where *P* – the proportion of words in the response that are present in the reference. It is defined as the ratio of the number of common words to the number of words in the response;

R – the proportion of words in the reference response that are present in the model response. It is defined as the ratio of the number of common words to the number of words in the reference response;

β – a coefficient that is the ratio of *P* to *R*.

Latency – the average response generation time:

$$Latency = \frac{T}{N},$$

where *T* – the sum of all response times.

Failure Rate – the proportion of requests for which the model did not provide any meaningful response:

$$Failure\ Rate = \frac{N_e}{N} 100\%,$$

where *Ne* – number of incorrect answers.

Expert evaluation was carried out on a scale from 1 to 5 according to the following criteria:

- quality of the answer (relevance, accuracy),
- clarity (simplicity of formulation),
- naturalness (is the answer like a human answer),
- trust in the answer (does the user believe that the information is correct).

The evaluation was carried out by 5 independent reviewers without technical training and 2 specialists in the field of customer support. Table 1 shows the values of the metrics for evaluating the results of the experiment. Figures 2–3 show histograms of automatic and expert indicators, respectively.

Table 1

Values of metrics for evaluating the models' performance

Metrics	Rule-Based	Retrieval-Based	LLM-Only	Hybrid model
Accuracy (%)	62.3	74.8	87.5	92.1
BLEU Score	0.45	0.61	0.72	0.78
ROUGE-L Score	0.51	0.65	0.76	0.81
Latency (c)	0.8	1.2	2.4	1.6
Failure Rate (%)	18.2	11.5	6.1	3.7
Answer quality (1–5)	3.2	3.9	4.4	4.7
Intelligibility (1–5)	3.1	3.8	4.5	4.8
Naturalness of language (1–5)	2.7	3.5	4.8	4.9
User trust (1–5)	2.9	3.6	4.3	4.6

Source: created by the author

The conducted experiment made it possible to obtain quantitative assessments of the quality of the hybrid model and three basic models of processing customer requests in chatbots. The hybrid model demonstrated the highest average BLEU Score among all participants in the experiment. This indicates a high similarity between the generated responses and reference texts by n-grams. In particular, the rule-based model had the

lowest BLEU, which was expected because of the limited number of predefined scenarios. According to the ROUGE-L indicator, the proposed hybrid model also outperformed other solutions, demonstrating a better ability to preserve the main content of reference responses.

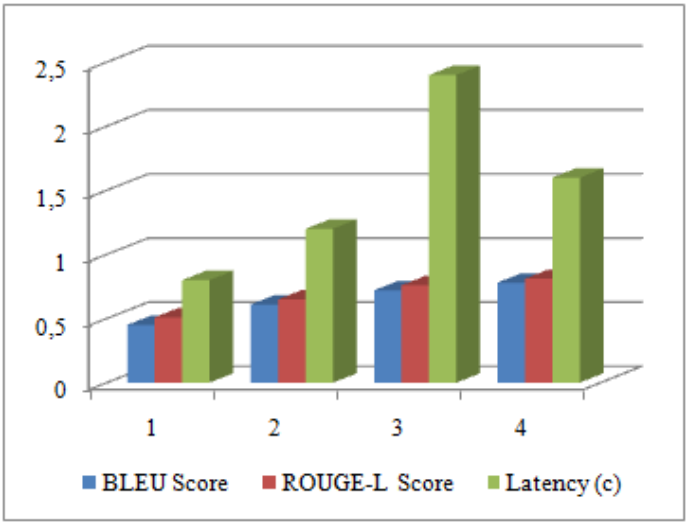

Fig. 2. Histogram of BLEU, ROUGE-L, Latency metrics for the studied models: 1 – Rule-Based, 2 – Retrieval-Based, 3 – LLM-Only, 4 – hybrid model

Source: created by the author

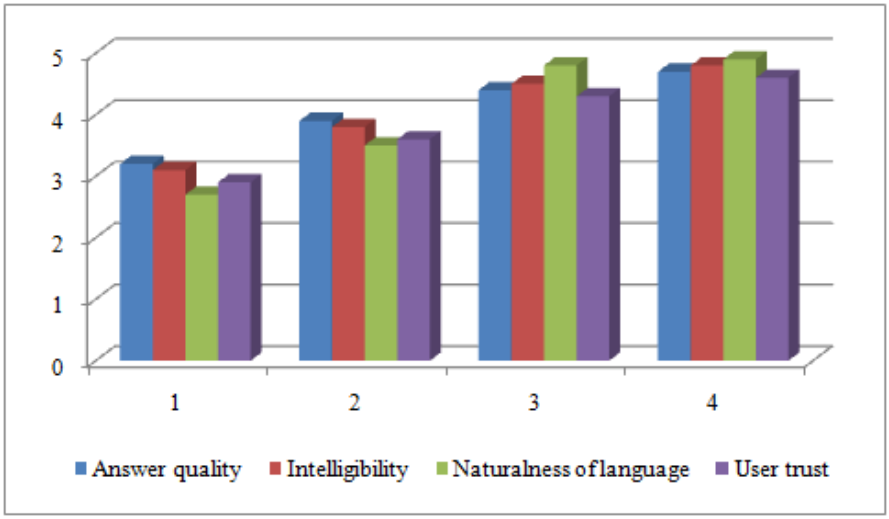

Fig. 3. Histogram of expert metrics for evaluating results for the studied models: 1 – Rule-Based, 2 – Retrieval-Based, 3 – LLM-Only, 4 – hybrid model

Source: created by the author

Fine-tuned LLM showed a medium result, while the rule-based approach significantly lost ground, probably because of the lack of adaptation to the specifics of the requests. The hybrid model received the highest scores for the Intelligibility criterion, which indicates the intuitiveness and ease of perception of its texts by users. Particularly large differences were observed between the hybrid model and the rule-based bot, which sometimes generated too formal or limited in content answers.

The naturalness of the hybrid model's answers was rated higher than that of other models, indicating a better imitation of the human communication style. The rule-based bot showed the lowest results because of rigid template constructions in the answers. The hybrid system ranked took first place in terms of the accuracy of the provided answers. This emphasizes its ability not only to create understandable and natural phrases, but also to adequately resolve users' information requests.

All results showed a low level of variance of estimates for the hybrid model, which indicates its stability on different subsets of requests. Integration of a specialized semantic pre-processor with generative capabilities of LLM significantly improves both formal metrics of correspondence (BLEU, ROUGE-L) and human impressions of interaction (Intelligibility, Naturalness).

Basic approaches have certain advantages in specific cases, but lose efficiency when the variety of user requests increases. The results of the conducted research have high practical significance for a wide range of applications in the field of automation of processing of client requests. The proposed hybrid model, which combines pre-processing of requests using semantic classifiers with the generative power of LLMs, demonstrates a significant improvement in the quality of interaction with users. An example of a request and responses is given below:

Example of user request 1:

"Good afternoon! I haven't received my bills in the mail for two months. What's the problem?"

Answer of the Rule-Based template system:

"You can download the bill in your personal account at the link…"

LLM answer:

"Sorry for the inconvenience. Try checking your spam folder or updating your e-mail in your personal account."

Hybrid model answer:

"There was probably a failure in sending bills. Check the e-mail in the user account or contact the operator. If the e-mail is correct, I will resend the bill."

Example of user request 2:

"I can no longer pay for the Internet at my tariff. It is now expensive for me. How can I suspend payment for the service?"

Answer of the Rule-Based template system: "Payment information is available in the Tariffs section of the website."

LLM answer:

"I understand that you are in a difficult situation right now. You may want to contact our support team to review your plan."

Hybrid model answer:

"We are working with a social support programme. Call 0-800-xxx or fill out a form to temporarily suspend payments."

Both examples demonstrate the advantage of the hybrid model – contextual sensitivity and a specific, actionable response that goes beyond a template approach. The rule-based system provides general information, LLM shows empathy, but does not always offer a practical solution. Instead, the hybrid model combines an emotional response with a real next step, which is important in a service environment.

The proposed hybrid model will contribute to:

– Increasing customer satisfaction due to high indicators of clarity, naturalness, and accuracy of answers. The system generates responses that are stylistically and emotionally close to human communication, which has a positive effect on the perception of the service by customers.

– Reducing the workload on contact centre operators. More accurate and relevant automatic answers significantly reduce the number of cases of escalation of appeals to people.

– Adaptability to new scenarios. Unlike classic rule-based systems, the hybrid model can quickly integrate new types of requests without significant costs for rewriting scripts.

– Reducing response time and increasing service efficiency, which directly affects the companies' business indicators.

The hybrid approach to processing appeals combines semantic classification with a generative language model. This is an improvement on existing architectures designed to achieve a better balance between the structuring and flexibility of answers. The proposed methodology for assessing the quality of the request processing system with a combination of automatic metrics (BLEU, ROUGE-L) and subjective human assessments (Intelligibility, Naturalness, Accuracy) provides a comprehensive approach to analysing effectiveness. Clarification of the characteristics of the BLEU and ROUGE metrics specifically for short dialogic responses, which was not always taken into account in similar studies before.

So, the study contributes both to the theory of LLM integration into application service systems and to practical methods for building more flexible and effective new generation chatbots. Thanks to its architecture, the proposed model can be applied in various areas, such as banking, e-commerce, technical support, healthcare, government services, and others. It is effective in those industries where high-quality customer service is required 24/7. The conducted research opens up a number of promising areas for further improvement. For

ISSN 2786-5460 (Print), ISSN 2786-5479 (Online)

example, the introduction of active learning methods, which will allow the system to independently identify the most problematic types of requests and request human assistance for clarification. This will increase accuracy without the need for massive manual data labelling.

The hybrid architecture can also be further expanded by personalizing responses for a specific user, using his history of requests and behavioural patterns. In particular, the next step may be to integrate the processing of not only text, but also voice, visual or structured customer requests, which will significantly expand the capabilities of the system. Besides, the analysis of the security of hybrid systems will be an important direction: studying their behaviour in response to aggressive, manipulative, or unethical user requests. It is also worth researching options for using less resource-intensive models or knowledge distillation methods to build compact, energy-efficient solutions.

Research limitations. Despite the positive results, the study has a number of limitations that should be taken into account when interpreting the obtained findings. The models were tested on a limited corpus of requests, which, although covering the main types of requests, do not fully reflect the full diversity of speech behaviour of real users. It should also be noted that LLM was run in a relatively stable, no-load environment during the experiment, while performance may be affected by external API delays or resource constraints in real-world applications. Furthermore, the hybrid model requires maintaining up-to-date data (contacts, policies, programme terms), which poses scalability and maintenance challenges. Assessing the clarity and relevance of responses also remains somewhat subjective.

Implementation prospects. The integration of LLMs into customer service systems opens up new opportunities for business and the public sector. In technical support services, LLMs can automatically respond to user requests, reducing the load on operators and response times. Online stores can use LLMs for personalized recommendations, order processing, and resolving customer issues in real time. Automating responses to typical citizen requests, such as processing documents or receiving social benefits, can increase the efficiency and accessibility of government services.

Implementation barriers. Despite significant advantages, there are certain challenges:

1. Integration with existing systems. Many organizations use legacy systems with which it is difficult to integrate modern LLMs.

2. The need for moderation. LLMs can generate incorrect or unacceptable answers, so a mechanism for checking and moderating content is required.

3. Updating knowledge bases. It is necessary to regularly update the databases from which LLM receives data to ensure that the answers are up-to-date.

Conclusions. A new hybrid architecture for automatic customer request processing was developed and tested in the study, which significantly improves the quality of interaction with the user compared to existing approaches. The proposed model demonstrates high results according to the selected metrics for assessing the quality of responses and confirms the effectiveness of combining semantic routing of requests with the capabilities of generative language models. The results of the study have significant practical value for implementation in various areas of business and open up new opportunities for the further development of intelligent customer service systems. The proposed approaches also form the basis for further research aimed at personalization, multimodal request processing, increasing the dialogue system security, and optimizing resource costs.

Bibliography:

1. Asai A., Min S., Zhong Z., Chen D. Retrieval-based language models and applications. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts), July 2023. P. 41–46. URL: https://aclanthology.org/2023.acl-tutorials.6/ (date of access: 1.05.2025).

2. Hong M., Song Y., Jiang D., Wang L., Guo Z., Zhang C. J. Expanding chatbot knowledge in customer service: Context-aware similar question generation using large language models. arXiv preprint arXiv:2410.12444. URL: https://arxiv.org/abs/2410.12444 (date of access: 1.05.2025).

3. Jo S., Seo J. ProxyLLM: LLM-driven framework for customer support through text-style transfer. arXiv preprint arXiv:2412.09916. URL: https://arxiv.org/abs/2412.09916 (date of access: 1.05.2025).

4. Kaddour J., Harris J., Mozes M., Bradley H., Raileanu R., McHardy R. Challenges and applications of large language models. arXiv preprint arXiv:2307.10169. URL: https://arxiv.org/abs/2307.10169 (date of access: 1.05.2025).

5. Kruk F., Herath S., Choudhury P. BanglAssist: A Bengali-English generative AI chatbot for code-switching and dialect-handling in customer service. arXiv preprint arXiv:2503.22283. URL: https://arxiv.org/abs/2503.22283 (date of access: 1.05.2025).

6. Li X., Gao M., Zhang Z., Yue C., Hu H. Rule-based data selection for large language models. arXiv preprint arXiv:2410.04715. URL: https://arxiv.org/abs/2410.04715 (date of access: 1.05.2025).

7. Marr B. The amazing ways Duolingo is using AI and GPT-4. URL: https://bernardmarr.com/the-amazing-ways-duolingo-is-using-ai-and-gpt-4/?utm_source=chatgpt.com (date of access: 1.05.2025).

8. Naveed H., Khan A. U., Qiu S., Saqib M., Anwar S., Usman M. та ін. A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435. URL: https://arxiv.org/abs/2307.06435 (date of access: 1.05.2025).

9. Nerella S., Bandyopadhyay S., Zhang J., Contreras M., Siegel S., Bumin A. та ін. Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, 2024. Vol. 106. Art. 102900. DOI: https://doi.org/10.1016/j.artmed.2024.102900.

10. Rüdel T., Leidner J. L. Control in hybrid chatbots. arXiv preprint arXiv:2311.11701. URL: https://arxiv.org/abs/2311.11701 (date of access: 1.05.2025).

11. Wulf J., Meierhofer J. Utilizing large language models for automating technical customer support. arXiv preprint arXiv:2406.01407. URL: https://arxiv.org/abs/2406.01407 (date of access: 1.05.2025).

12. Xu Z., Cruz M. J., Guevara M., Wang T., Deshpande M., Wang X., Li Z. Retrieval-augmented generation with knowledge graphs for customer service question answering. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 2024. P. 2905–2909. DOI: https://doi.org/10.1145/3626772.3661370.

13. Yao Y., Wang P., Tian B., Cheng S., Li Z., Deng S. та ін. Editing large language models: Problems, methods, and opportunities. arXiv preprint arXiv:2305.13172. URL: https://arxiv.org/abs/2305.13172 (date of access: 1.05.2025).

Дата надходження статті: 23.06.2025
Дата прийняття статті: 04.07.2025
Опубліковано: 23.09.2025