Scientific Journal of the Ternopil National Technical University
2026, vol. 121, no. 1, pp. 37–45,
ISSN: 2522-4433 Print, ISSN: 2522-4441 Online
https://doi.org/10.33108/visnyk_tntu2026.01.037

Submitted: 2025-12-28 | Revised: 2026-03-14 | Accepted: 2026-03-24

Development of an intellectualized program system based on rag-architecture

Daria Zavtrak [1], Valentyn Khychan[2], Yeva Dubanych[3], Bohdan Dobrotvor[4], Yaroslav Herasimchuk[5], Borys Bessarab[6], Artem Mandziy[7], Oleh Pastukh[8]

[1] Ternopil Ivan Puluj National Technical University, Ukraine, zavdariag@gmail.com

[2] Ternopil Ivan Puluj National Technical University, Ukraine, valikkican5@gmail.com

[3] Ternopil Ivan Puluj National Technical University, Ukraine, evadubanich8@gmail.com

[4] Ternopil Ivan Puluj National Technical University, Ukraine, dobrotvorbo@gmail.com

[5] Ternopil Ivan Puluj National Technical University, Ukraine, yarosmen026@gmail.com

[6] Ternopil Ivan Puluj National Technical University, Ukraine, borys_bessarab0608@tntu.edu.ua

[7] Ternopil Ivan Puluj National Technical University, Ukraine, 2005art.man@gmail.com

[8] Ternopil Ivan Puluj National Technical University, Ukraine, oleg.pastuh@gmail.com

Abstract: *The article presents an approach in the creation of an intellectualized dialogue system for automation of communication with university applicants during admission campaigns. Retrieval-Augmented Generation (RAG) architecture, which combines search of relevant information with generation of a response using a Large Language Model, is the basis of the development. Fine-tuning of the GPT-2 model with regulatory documentation of a higher educational facility was used to provide support for Ukrainian language. A hybrid mechanism of response generation, which combines fragments extracted from a knowledge base with segments generated by the language model, was proposed. An architecture, which ensures the relevance, accuracy and completeness of data, as well as lowering the workload on university employees, was realized. The quality of the system was rated using both objective (response speed, BERTScore, relevance) and subjective metrics (completeness, convenience, search flexibility), which allowed to record the efficiency of the approach. Presented instrument demonstrates the perspective of the usage of RAG approaches in applied tasks of the educational sphere, especially in support systems for admission committees.*

Keywords: *Artificial Intelligence, chatbot, dialogue system, Language Model, Large Language Model, generation, prompt, RAG, fine-tuning, accuracy, BERTScore.*

1. INTRODUCTION

Usage of artificial intelligence by employees of state institutions slowly becomes a regular practice, especially in the sphere of education. According to the research conducted by Microsoft and LinkedIn, approximately 75% of employees (Fig. 1) involve artificial intelligence in their professional activity [1]. In the context of these changes the need for optimization of certain aspects of work of public state employees, especially the process of admission campaigns for higher educational facilities, emerges.

Annually during the period of admission campaigns, the amount of responsibilities of university employees significantly increases, especially because of the need to provide

consultations and information to university applicants regarding the admission peculiarities. The direct communication with university applicants does not always ensure the proper level of awareness, meanwhile independent search of information in regulatory documentation often is a long and complex process. The use of artificial intelligence for optimization of this process will fulfil the informational need of university applicants and greatly reduce the amount of work for university administrations.

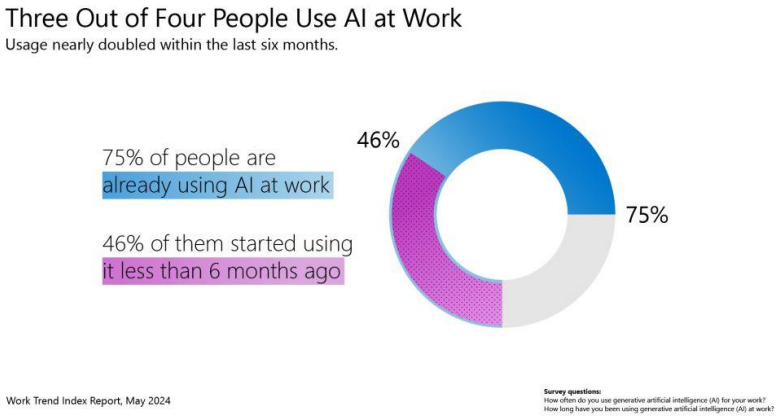

Figure 1. Microsoft and LinkedIn research infographics [1].

The development and training of our own artificial intelligence model is an effective solution that will guarantee the integrity and completeness of data. The research of new ways of answer generation is an important part of improving the quality of the interaction. The use of a dialogue system with the application of Retrieval-Augmented Generation (RAG)[2] allows to automate the communication with the user, decreasing the workload of admission committees, in the context of high competition for university applicants. Due to this approach the system forms responses based on relevant fragments of internal documentation and generated text, which significantly increases the quality of informing. A dialogue system would provide a convenient usage for university applicants. Maintenance and fine-tuning of a model is a significantly less time-consuming process than manual processing of all of applicants' prompts. This will ensure the accessibility of information at any time of the day, as well as decrease the service time.

2. LITERATURE OVERVIEW

The objective of literature overview is to inspect the subject area as comprehensively as possible, examine the already existing works of other scientists: analyze, generalize, compare the works among themselves and highlight what is already researched, and what is still undiscovered. In turn, this will involve analyzing the utilized approaches, methodologies and obtained results, highlighting the strengths and weaknesses that will give us a basis for formulating our contribution with this work to the research of the topic. Selection of sources was formed by established criteria to conduct analysis based on authoritative, relevant and approved data. Publications that are no older than five years, which had either identical or extremely similar subject to the one discussed in this work, were taken into account. Approaches used by other researchers were key factors in the selection as well.

In article [3], the authors developed a chatbot with the goal of making the work of university admission committee employees easier. With the utilization of Deep Learning models that were already integrated into the Rasa framework, they developed a model that can distinguish more than fifty types of questions entered by a user, having a 97,1% accuracy on

test sets. As the authors noted, there are two main types of chatbots with Deep Learning: retrieval-based bots and generative bots. In the article, a retrieval-based model was used for the construction of the NEU-chat. But because the bot specializes in retrieving information for documents and not generation, i.e., is different by giving an answer not based on current and previous messages from a user, this deprives it from the possible level of convenience of use, since it inherently doesn't utilize the natural human language.

In article [4] the authors improve on the already developed RAG mechanism that in turn combines mechanisms based on information retrieval with generative models to avoid the existing difficulties in utilization and implementation of the applicant consultation system. The Unified RAG (URAG) framework is utilized for improving light LLM's for chatbots for admission campaigns at universities. URAG combines the reliability of systems based on rules with adaptiveness of RAG, creating a two-level approach. The first level uses a complex system of frequently asked questions (FAQ) for providing accurate answers for typical questions, especially if they concern important or critically important information. If an answer was not found in the FAQ system, the second level retrieves corresponding documents from an expanded database and generates a repeated answer using the LLM. Based on the presented results, this approach shows high effectiveness, because it is better than the mentioned commercial chatbot models in a number of criteria, which testifies to significant perspectives of implementing similar mechanisms to create a system for consultations and more.

In the article [5] the authors, similar to the former paper, use the Rasa frameworks to create and customize their language model, which will be utilized for automating the process of proving answers to questions about admission to the local university. The interaction between the bot and the user includes the processing of the message, identification of the intent and entities, a choice of an action with help from policies and provision of an answer. The DIET model, which encompasses tokenization, creation of marks, recognition of entities and classification of intent, is used for NLU. The authors experimented with six variants of conveyors, changing methods of tokenization (VNCoreNLP, spaces) and vectorization (regular expressions, CountVector, BERT, PhoBERT) to find the most effective configuration for the functioning of the bot. It is worth pointing out that specific data preparation which includes admission data, questions and answers to them are key in this work. The NLU component is utilized for intent classification and entity extraction, which is a typical approach in retrieval chatbots, where the answers are dependent on the formerly defined templates and scenarios.

In article [6] the development of a university support system is described, which is based on generative artificial intelligence (GenAI-USS) that uses modified RAG architecture to increase effectiveness of LLMs with focus on step-by-step transparency. The goal is to ensure flexibility, transparency and accuracy of answers to the prompts using the data from the university websites and knowledge. The RAG architecture allows the utilization of the modular principal with the choice of appropriate hints (prompts) to increase accuracy. One of the key elements is integration of relevant information in real time from the formerly defined sources that creates a dynamic knowledge base, adapted for the subject area. The other important aspect is the transparency of data processing on all stages: from data and saving to coding, testing and interacting with the chatbot. The testing module allows to decrease hallucination, increase the answer accuracy and ensure their relevance, which leads to the creation of transparent, flexible and quality generative AI-system with RAG support.

Thereby, from the experience of the research already conducted by other scientists, having analyzed all of the requirements, priorities and capabilities for realization, in this work we have described a system, which utilizes RAG with some modifications, aimed at improving the user experience by increasing the speed and accuracy of the results for the given prompts. Having conducted approach hybridization, by combining results of semantic retrieval with fragments generated by the language model in the final answer, we were able to achieve the

results presented in the corresponding chapter of this work. Another neural network is responsible for the quality generalization of the final answer, which receives the corresponding prompts through an application programming interface (API). The key aspect that is worth mentioning, is that the fine-tuning, retrieval and generation are realized by language models using Ukrainian language, which means that the semantic basis is smaller than in models, trained on corpuses of documents in English language, which will have its influence on the results later.

3. DEVELOPMENT OF THE SYSTEM

The CPT-2 ukr architecture [7] adapted by fine-tuning [8] with the use of a corpus of regulatory and informative documents of the university was utilized for the development of the system that will meet all the requirements and get provided with enough data for the analysis of the end results, so that support of Ukrainian language, will be ensured. Chatbots based on Large Language Models require fine-tuning. Because of that, before the full realization and implementation of the chatbot it is suggested to overview a set of technologies that will ensure the high accuracy of answers, effectiveness of learning and scalability.

The basis of the system are the Transfer Learning and Fine-tuning of transformer model technologies. They provide the usage of a previously trained language model and its fine-tuning [9]. The advantage of using specifically this approach is saving resources of the computing machine by avoiding the need to fully train the model from the start, because it already has a deep semantic basis. Fine-tuning improves the accuracy in highly specific prompts as well, which ensures the decrease of the risk of model «hallucinations» in the context of a domain.

Data that is utilized for training, requires preparation in advance. Firstly, files in the .txt format are loaded into the system. Big text is being split into parts according to the limitations of the model with the use of Data Chunking technology. A data set from the list of text chunks is created, tokenization and creation of necessary marks occurs [10]. The processed data is used for fine-tuning of the model.

Additionally, the technology for redistribution of the tasks among available processors was applied for the optimization of the processes and decreasing the load on the computing machine [11].

Autoregressive Text Generation [12] was used for the realization of the interaction with a user in a dialogue regime. This ensures the convenience and intuitive interface. The system is built on RAG principle – at first, a search of three relevant fragments from the corpus of documents based on the user prompt occurs, after which a text fragment is generated additionally. The final answer forms as a generalization between the retrieved and the generated fragments that allows to reach a balance between accuracy and completeness. This structure is shown on Figure 2. For the generalization of the final answer another neural network (Gemini-2.0), to which a corresponding request is sent with the use of API, was utilized.

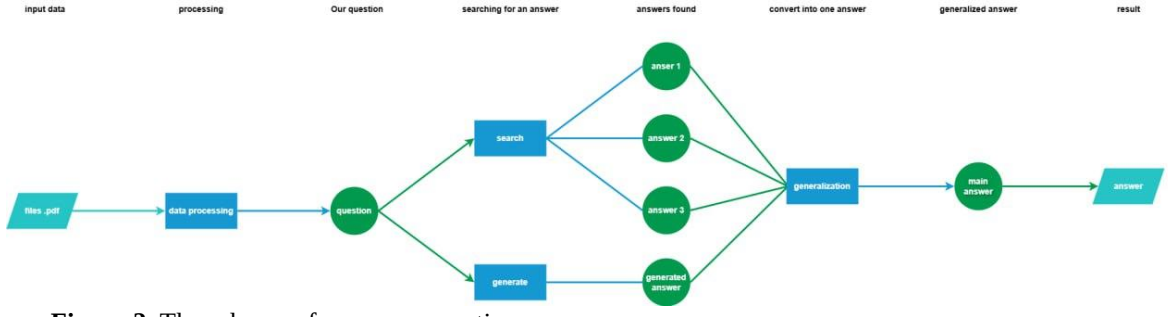

Figure 2. The scheme of answer generation.

The research of the new answer architecture is a key step in the development of the system. The optimization of the answer generation mechanism specifically determines the level of relevance, accuracy and completeness that the user receives.

4. RESULTS

Methods of objective and subjective rating of the quality by users were utilized to demonstrate the results. The end analysis and conclusions were formed on the basis of those ratings. The metrics presented below were utilized for numerical evaluation of the processing of user prompts by the developed model.

Objective rating metrics:

- The speed of receiving the answer.
- Answer accuracy (BERTScore) [14].
- Data relevance.

Subjective rating metrics:

- Information completeness.
- Convenience of use.
- Accessibility and ease of understanding of information.
- Search flexibility.

There is a need to define each of the used metrics for conducting the result analysis. In this context some of the used standard metrics may be utilized in partially changed ways [15].

The speed of receiving the answer – numerical metric that identifies time (in seconds) from the sending of the prompt to the receiving of the full answer.

Answer accuracy – percentage rating of factual accuracy, the correspondence of the answer to the prompt.

Data relevance – quality assessment of relevance to the topic of the prompt, its context and intention of the user.

Criteria of relevance:

- Semantic relevance of the prompt;
- The answer has to correspond to the content and intention of the user, even if the wording of the prompt is generalized or incorrect.
- The accuracy of retrieval from the base of knowledge.
- The system has to utilize specifically the fragments of the documents that directly relate to the prompt, without excessive generation or free interpretation of the content.
- Avoidance of the informational noise.

The answer cannot contain redundant or not relevant to the prompt facts.

Information completeness – subjective rating of the expected volume of the prompt answer, considering accuracy and relevance.

Convenience of use – subjective rating of the benefit of the use by the user.

Accessibility and ease of understanding of information – subjective rating of the ease of understanding, consistency and adaptability of the received answer by the user.

Search flexibility – subjective rating of the ability of the system to retrieve relevant information for the prompts, formulated using synonyms, rephrases and diverse language constructs.

Having collected the statistical data of rating results of the performance of the language model, the following results were received: the objective metrics rating is, on average, 7,98; the rating differentiates from the subjective rating by 0,13, which is 8,12. Percentage ratings of accuracy and time were converted into the decimal scale and averaging the criteria of relevance (the mean of all relevance criteria) were used for the calculation of an objective value.

The average ratings of the test group consisting of 10 people (objective rating was taken into account) for a number of received answers is demonstrated on Figure 3. Altogether there were 400 ratings received for the analysis. On the *y* axis the rating itself is illustrated, while the number of the prompts is on the *x* axis. From the data, shown on the presented graphic, we can calculate the standard error with the use of the formula:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

For this we calculated the standard deviation in advance and it is equal to 0,48. The standard error, in turn, is 1,94% in the limit of the sample.

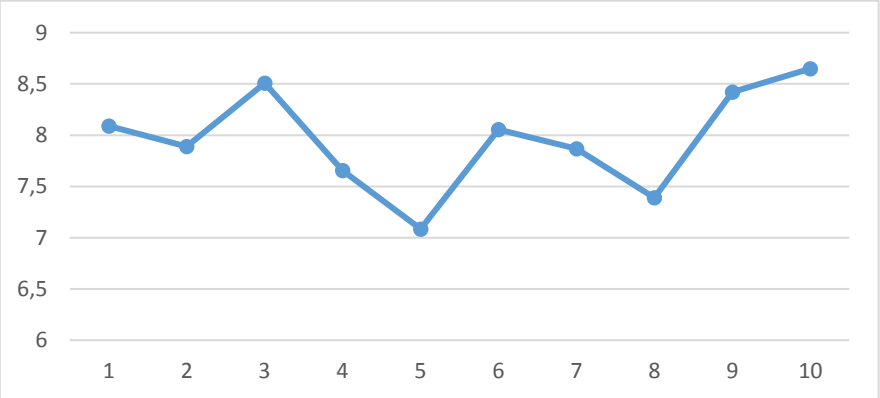

Figure 3. Average rating values graphic.

We can affirm the consistency of the sample, i.e., about the objectiveness of test group rating, because the margin of rating error is mild and the standard deviation is miniscule.

In the process of metric evaluation, the answer generation time data with identical prompts but different answer generation architecture were measured (Fig. 4–5).

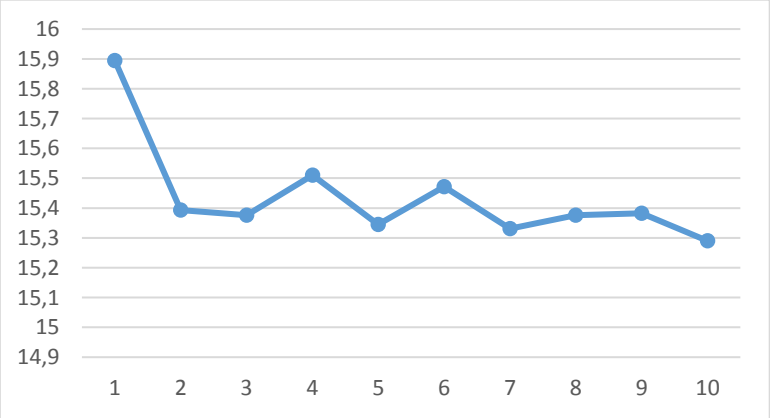

Figure 4. Average generation time graphic (standard GPT-2 ukr architecture).

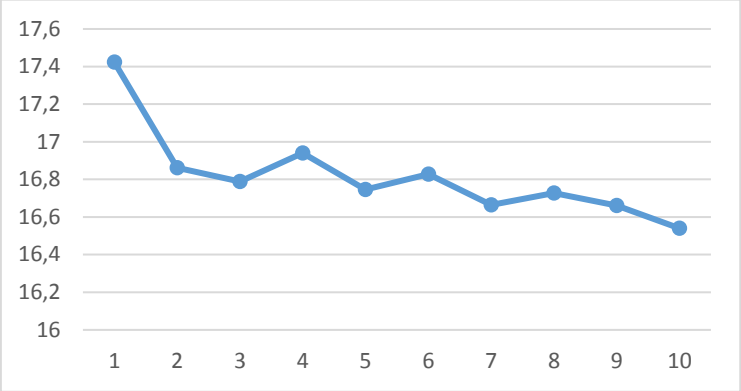

Figure 5. Average generation time graphic (presented model).

A similar dynamic is noticeable, because the answers were generated simultaneously. We get the following graphic (Fig. 6) after finding the difference between the metric measurements.

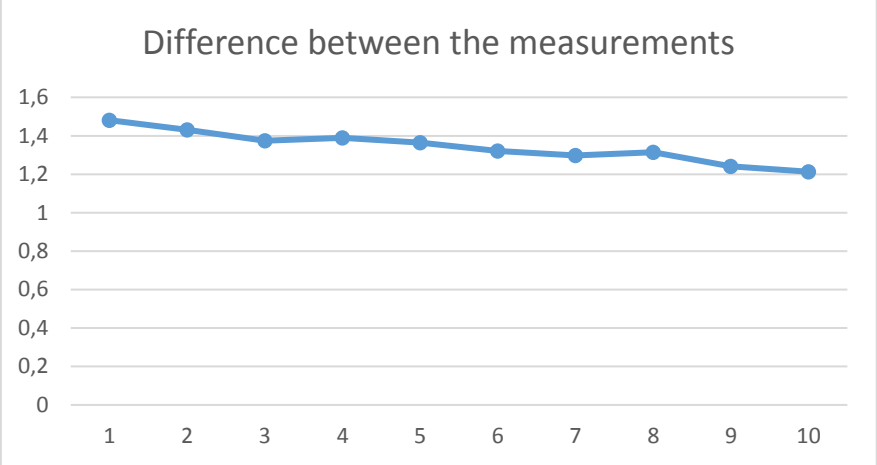

Figure 6. Difference between the measurements

It is apparent that the model presented by us has a somewhat slower answering speed. The difference between the generation is, on average, 1,3 seconds, which is not critical for the user experience. In turn, the utilized architecture based on RAG improves the value of the answer accuracy by 9% on average, which not only increases the answer quality, but also insignificantly improves the user experience.

5. ANALYSIS OF THE RESULTS

The insignificant difference between the average values of subjective and objective metric ratings can be evident from reconciliation of the answer quality from the technical perspective and the subjective perception of the quality by the users, with an insignificant error towards the subjective rating. This testifies to the balance between the user expectations and the current quality of the model.

Due to the reconciliation, the inductive reasoning method can be applied for forecasting the quality of further research, in particular, to extrapolate the conclusions to the wider audience. In general, this evaluation method requires a larger sample size for a more accurate forecast, as well as defining the correlation and possible multicollinearity of the ratings of the different metrics that were used in the research. With a sample of 400 rating, it was not possible to accurately establish relationships between the indicators of the metrics of different

objectivity nature. At this stage, the analysis of the correlation coefficient cannot be considered statistically significant because of the low value of the Pearson correlation coefficient.

The insignificant decrease of the answer speed is compensated by the notable improvement of the answer quality by 9%. This improvement confirms the effectiveness of the RAG usage for the language model architecture, the received results were expected and are consistent with the results of the researches reviewed earlier. So, we can assert the effectiveness of the use of the presented architecture of answer construction.

6. CONCLUSIONS

The developed dialogue system applies the RAG architecture. The model was fine-tuned with the utilization of the corpus of university documents in Ukrainian. This guarantees high accuracy and relevance of the answers to the prompts of the applicants, the use of the reliable data from the internal sources and decreasing the model's «hallucinations».

The system demonstrates the improvement of answer accuracy by 9% compared to the standard architecture, which guarantees receiving accurate and full information by the applicants. This increases their general satisfaction and trust in the system. Despite the insignificant increase in answer generation time (approximately 1,3 seconds), the answer quality remains high, which testifies to the optimal balance between speed and quality. The users are ready to wait slightly longer to receive a more accurate and informative answer, which improves the general user experience.

The ratings of the system quality (both objective and subjective) are high and reconciled. This confirms that the system is not only technically functional, but also convenient and understandable for the end users, effectively fulfilling their informational needs.

The implementation of this system automatizes the process of consulting. This significantly decreases the workload for the employees of the admission commissions during peak load of the admission campaign and ensures the accessibility of the information for the applicants at any time of the day.

REFERENCES

[1] Three out of four people use AI at work. Microsoft. (2024). https://news.microsoft.com/annual-wti-2024/.

[2] J. Swacha, M. Gracel, Retrieval-augmented generation (RAG) chatbots for education: A survey of applications, Applied Sciences, 15 (2025), 4234. https://doi.org/10.3390/app15084234.

[3] T.T. Nguyen, et al., NEU-chatbot: Chatbot for admission of National Economics University, Computers and Education: Artificial Intelligence. 2 (2021) 100036. https://doi.org/10.1016/j.caeai.2021.100036.

[4] N. Chidipothu, et al., Improving large language model (LLM) performance with retrieval-augmented generation (RAG): Development of a transparent generative AI university support system for educational purposes, Journal of Big Data and Artificial Intelligence, 3 (2025), 1. https://doi.org/10.54116/jbdai.v3i1.50.

[5] M.-T. Nguyen, et al., Building a chatbot for supporting the admission of universities, in: 2021 13th International Conference on Knowledge and Systems Engineering (KSE), Bangkok, Thailand, 2021. https://doi.org/10.1109/KSE53942.2021.9648677.

[6] L.S.T. Nguyen, T.T. Quan, URAG: Implementing a unified hybrid RAG for precise answers in university admission chatbots – A case study at HCMUT, Communications in Computer and Information Science. (2025) 82–93. https://doi.org/10.1007/978-981-96-4285-4_7.

[7] Malteos, gpt2-uk, Hugging Face. https://huggingface.co/malteos/gpt2-uk.

[8] E. Latif, X. Zhai, Fine-tuning ChatGPT for automatic scoring, Computers and Education: Artificial Intelligence. 6 (2024) 100210. https://doi.org/10.1016/j.caeai.2024.100210.

[9] N. Ding, et al., Enhancing chat language models by scaling high-quality instructional conversations, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 2023. https://doi.org/10.18653/v1/2023.emnlp-main.183.

[10] C.W. Schmidt, et al., Tokenization is more than compression, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 2024, pp. 678–702. https://doi.org/10.18653/v1/2024.emnlp-main.40.

[11] D. Gyawali, Comparative analysis of CPU and GPU profiling for deep learning models, arXiv. (2023). https://arxiv.org/abs/2309.02521.

[12] Z. Ul Abideen, Autoregressive models for natural language processing, Medium. https://medium.com/@zaiinn440/autoregressive-models-for-natural-language-processing-b95e5f933e1f.

[13] M. Prytula, O. Sinkevych, I. Olenych, Comparison of zero-shot approach and retrieval-augmented generation for analyzing the tone of comments in the Ukrainian language, Electronics and Information Technologies. 28 (2024). https://doi.org/10.30970/eli.28.1.

[14] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, arXiv. (2019). https://arxiv.org/abs/1904.09675.

[15] V. Yatsyshyn, et al., Technology of relational database management systems performance evaluation during computer systems design, Scientific Journal of the Ternopil National Technical University. 109 (2023) 59–64. https://doi.org/10.33108/visnyk_tntu2023.01.054.