**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ХАРЧОВИХ ТЕХНОЛОГІЙ**

**WARSAW UNIVERSITY OF TECHNOLOGY (М. ВАРШАВА, ПОЛЬЩА)**

**INTERNATIONAL INFORMATION TECHNOLOGY UNIVERSITY (М. АЛМАТИ, КАЗАХСТАН)**

**ІНСТИТУТ ПРОБЛЕМ МАТЕМАТИЧНИХ МАШИН І СИСТЕМ НАЦІОНАЛЬНОЇ АКАДЕМІЇ НАУК УКРАЇНИ**

**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»**

**ЧЕРКАСЬКИЙ ДЕРЖАВНИЙ ТЕХНОЛОГІЧНИЙ УНІВЕРСИТЕТ**

**НАЦІОНАЛЬНИЙ ТРАНСПОРТНИЙ УНІВЕРСИТЕТ**

**ТАВРІЙСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ В. І. ВЕРНАДСЬКОГО**

AI&IT

Третя міжнародна
науково-практична конференція

# «Штучний інтелект та інформаційні технології»

1–2 червня 2026 р.

Київ НУХТ 2026

УДК 004

Наукові праці Третьої міжнар. наук.-практ. конф. «Штучний інтелект та інформаційні технології» (AIIT-2026), 1–2 червня 2026 р. (Київ, Україна). К. : НУХТ, 2026. 420 с.

У працях конференції наведено доповіді за напрямами:

- тенденції та досягнення в розробленні й застосуванні методів і практичних інструментів штучного інтелекту;
- інтелектуальні системи управління та аналізу даних;
- використання інформаційних технологій та штучного інтелекту в освіті;
- кіберзахист критичної інформаційної інфраструктури;
- використання інтернет-речей у науці й виробництві;
- математичне моделювання складних об’єктів.

Праці конференції будуть корисні науковим та інженерно-технічним працівникам, викладачам і здобувачам ЗВО та всім, хто цікавиться сучасними інформаційними системами та телекомунікаційними технологіями.

**Подано в авторській редакції.**

*Автори матеріалів несуть повну відповідальність за достовірність наведеної інформації та відповідність матеріалів нормам законодавства, моралі й етики.*

**ISBN 978-966-612-426-8** 

UDC 004

Proceedings of the 3rd international scientific and practical conference «Artificial Intelligence & Information Technology» (AIIT-2026), June 1–2, 2026 (Kyiv, Ukraine). Kyiv : NUFT, 2026. 420 p.

The proceedings contain papers on the following topics:

- trends and achievements in the development and application of methods and practical tools of artificial intelligence;
- intelligent systems for data management and analysis;
- using information technology and artificial intelligence in education;
- cybersecurity of critical information infrastructure;
- using the Internet of Things (IoT) in science and production;
- mathematical modeling of complex objects.

The collection will be useful to scientists, researchers, professors, students, and everyone interested in modern information technology and artificial intelligence.

**Submitted in the authors’ edition.**

*The authors are fully responsible for the accuracy of provided information, as well as for the papers’ compliance with the laws, morals and ethics.*

**ISBN 978-966-612-426-8** 

UDC 004

# INTEGRATION OF VECTOR DATABASES INTO DECISION SUPPORT SYSTEMS FOR AUTOMATING THE WORK OF AUTOMOTIVE SERVICE TECHNICIANS USING RETRIEVAL-AUGMENTED GENERATION

**Kotvytska A., Seidykh O.**
*National University of Food Technologies, Kyiv, Ukraine*
*E-mail: kotvyckaaa@nuft.edu.ua*

*The automotive service industry is characterized by a high dependence on technical documentation, including OEM manuals, repair instructions, diagnostic trouble trees, and component specifications. The primary challenge is knowledge fragmentation and the difficulty of gaining rapid access to relevant information during the diagnostic process. This paper discusses an approach to developing a decision support system for automotive service technicians based on Retrieval-Augmented Generation (RAG) utilizing vector databases for the semantic search of technical information.*

The proposed architecture is based on a combination of information retrieval mechanisms and generative models. At the core of the RAG approach lies the principle of augmenting a large language model's (LLM) response with external knowledge sources, which enhances accuracy and reduces erroneous or fabricated answers (hallucinations) [1].

To validate the efficacy of the proposed approach, it is crucial to analyze RAG against alternative methods of domain adaptation, specifically the fine-tuning of large language models on technical manuals. While fine-tuning embeds domain knowledge directly into the model’s parametric weights, it suffers from three critical drawbacks in the automotive sector: high computational cost of frequent updates, vulnerability to catastrophic forgetting, and the inability to provide strict source citations for generated diagnostic steps. In contrast, the RAG architecture completely decouples the knowledge base from the language model. This allows for dynamic, real-time updates to the document registry (e.g., adding a newly released car model's manual) without retraining the underlying network, while ensuring absolute source traceability for the technician.

The technical implementation of the system involves multi-stage data processing. In the first phase, technical documentation from various automakers is collected in PDF, HTML, or structured database formats. Next, text segmentation into logical fragments (chunking) is performed, which preserves the contextual integrity of instructions and procedures.

Each fragment is converted into a vector representation (embedding) using neural models, specifically transformer architectures, which enable the encoding of the semantic meaning of text in a multidimensional space. A similar approach to constructing sentence embeddings is described in detail in the Sentence-BERT model [3].

The resulting vectors are indexed in specialized vector databases, such as FAISS or Qdrant. FAISS provides high-performance nearest neighbor search across

large vector datasets due to optimized indexing algorithms and the utilization of GPU computing [2].

Furthermore, the justification for using vector databases lies in overcoming search latency. A traditional exact search scales linearly (O(N)) with the number of document chunks, introducing unacceptable delays during real-time diagnostics. By leveraging Approximate Nearest Neighbor (ANN) search, the complexity is reduced to logarithmic time (O(log N)). This ensures that even when scaling the database to millions of documentation fragments, the system maintains sub-millisecond retrieval speeds with high accuracy, proving its operational viability for automotive services.

During system operation, a user query formulated in natural language (e.g., “unstable operation of the BMW N46 engine during cold start”) is also converted into a vector. Next, a search for the most semantically similar fragments in the vector database is executed based on cosine similarity or $L_2$ distance metrics.

The retrieved fragments of technical documentation form a context that is passed to the large language model. The model then generates a response in the form of a structured technical explanation or a step-by-step diagnostic algorithm. This approach eliminates the need for the model to possess direct "knowledge" of specific automotive systems, as it operates exclusively with up-to-date external data.

A key advantage of the RAG approach is a significant reduction in generative errors (hallucinations), which is a typical issue for large language models operating without an external context. By leveraging the retrieval component, the model is constrained to verified information from the knowledge base, thereby increasing the reliability of its responses [1].

In the context of automotive service, this is of critical importance, as diagnostic errors can lead to incorrect repairs, additional expenses, and damage to vehicle components. Therefore, the integration of semantic search significantly reduces the time required to access relevant instructions and enhances decision-making accuracy at the technician level. It is also worth noting that the use of vector databases ensures system scalability. Even when processing hundreds of thousands or millions of documentation fragments, search performance remains high due to Approximate Nearest Neighbor (ANN) search algorithms implemented in FAISS [2].

Thus, the proposed approach enables the transformation of traditional archives of technical documentation into a dynamic knowledge system that combines semantic search with the generative capabilities of language models. This establishes a foundation for next-generation intelligent assistants in the automotive industry, capable of functioning as interactive, real-time decision support systems.

## References


1. Lewis P. et al. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474.
2. Johnson J., Douze M., Jégou H. (2019) Billion-scale similarity search with GPUs, *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547.
3. Reimers N., Gurevych I. (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks, *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992.