Research on developing an offline intelligent system for scientific research data exploitation using large language models and retrieval-augmented generation

Nguyen Duc Hanh, Le Thi Thu Hong*

Institute of Information Technology and Electronics, Academy of Military Science and Technology, 17 Hoang Sam, Nghia Do, Hanoi, Vietnam.
*Corresponding author: lethithuhong1302@gmail.com
Received 9 Apr. 2026; Revised 18 Jun. 2026; Accepted 2 Jul. 2026; Published 15 Oct. 2026.
DOI: https://doi.org/10.54939/1859-1043.j.mst.114.2026.175-182

ABSTRACT

This paper presents an offline intelligent framework for exploiting scientific research management databases in secure internal network environments. The proposed system integrates open-source Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) to support natural language interaction with relational databases without requiring SQL expertise. The framework combines three core components: (1) a Text-to-SQL module that converts user queries into executable SQL statements, (2) a RAG mechanism that enriches query generation using database schemas, metadata, and domain knowledge, and (3) locally deployed open-source LLMs optimized for offline inference to ensure data privacy and operational independence. A benchmark dataset comprising approximately 500 natural language–SQL pairs was constructed for evaluation. Experimental results on Qwen3, Meta-Llama, and Gemma demonstrate that the proposed framework achieves reliable SQL generation, efficient inference, and secure offline deployment, thereby improving the accessibility and efficiency of scientific research data exploitation while satisfying stringent security requirements.

Keywords: Text-to-SQL; Retrieval-augmented generation; Large Language models; Scientific research management.

1. INTRODUCTION

Digital transformation has become a major driver of organizational modernization by integrating technologies such as Artificial Intelligence (AI), big data analytics, cloud computing, and the Internet of Things (IoT) to improve operational efficiency and decision-making [1]. As organizations accumulate large volumes of structured and unstructured data, effective information retrieval and knowledge exploitation have become increasingly challenging [2]. Scientific research management systems exemplify data-intensive environments containing information on research projects, personnel, publications, and funding within complex relational databases. Accessing these databases typically requires SQL proficiency and an understanding of database schemas, while complex queries often involve joins and aggregations. Consequently, non-technical users encounter significant barriers, and the semantic gap between natural language and database schemas further limits conventional query methods [3].

Recent advances in Large Language Models (LLMs) have enabled natural language interaction with relational databases through Text-to-SQL, reducing the need for SQL expertise and improving semantic query generation [4]. Despite these advances, practical deployment remains challenging because of complex database schemas, heterogeneous data structures, and domain-specific business rules. Accurate SQL generation requires effective schema understanding, semantic schema linking, and contextual reasoning, which continue to be major research challenges [5]. Furthermore, existing benchmark datasets, such as Spider, differ substantially from specialized enterprise databases in both schema complexity and domain terminology, limiting their applicability to real-world scientific research management systems [3, 6]. In addition, processing domain-specific Vietnamese queries poses notable linguistic and semantic challenges, where advanced deep learning architectures for Vietnamese text processing play a crucial role in capturing contextual features and categorization [7].

Retrieval-Augmented Generation (RAG) has become an effective solution for enhancing the reliability and contextual understanding of LLM-based systems by retrieving relevant external knowledge during inference, thereby reducing hallucinations and improving factual accuracy [6, 8]. Recent advances, including hybrid retrieval, metadata-aware search, semantic chunking, tabular text embeddings [13], and knowledge graph integration, further improve retrieval quality and contextual relevance [9, 10]. In Text-to-SQL applications, RAG provides access to database schemas, metadata, and example queries, effectively bridging the semantic gap between user intent and database structures. Meanwhile, strict security requirements in enterprise and government environments have encouraged the adoption of open-source LLMs deployed within private infrastructures, ensuring data privacy, regulatory compliance, and operational independence [11]. Moreover, instruction tuning and Reinforcement Learning from Human Feedback (RLHF) have demonstrated substantial improvements in the performance of locally deployed models for domain-specific applications [12].

Motivated by these challenges, this study proposes an offline intelligent framework for natural language interaction with scientific research management databases. The framework integrates Text-to-SQL generation, RAG, and locally deployed open-source LLMs to provide secure, accurate, and efficient data access in security-sensitive environments. Contextual retrieval enhances schema understanding and SQL generation, thereby narrowing the semantic gap between user queries and database structures. The remainder of this paper is organized as follows. Section 2 describes the proposed methodology and experimental setup. Section 3 presents the experimental results and discussion. Finally, Section 4 concludes the paper and outlines future research directions.

2. METHODOLOGY

2.1. Overview of proposed method

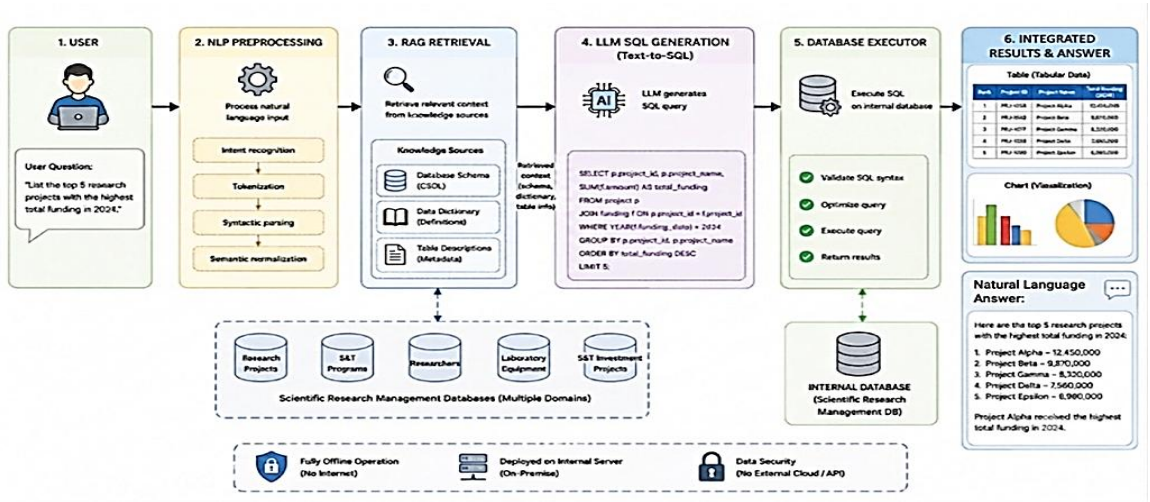

Figure 1. *The proposed intelligent system for scientific research data exploitation.*

This study proposes an offline intelligent framework for exploiting scientific research management databases through natural language interaction. The framework enables non-technical users to retrieve information, perform statistical analysis, and generate reports without requiring knowledge of SQL or database schemas. As illustrated in Figure 1, the system consists of six interconnected components. User queries are first processed by an NLP module for intent recognition and semantic analysis. A RAG module then retrieves relevant contextual information, including database schemas, metadata, and table descriptions, to enhance semantic understanding. The retrieved context is provided to a locally deployed LLM, which generates executable SQL

queries using a Text-to-SQL mechanism. These queries are validated and executed on the scientific research management database, which contains information on research projects, scientific personnel, laboratory equipment, and investment programs. Finally, the results are presented as tables, visualizations, and natural language summaries. The entire framework operates fully offline, ensuring data security, privacy, and independence from external cloud services.

2.2. Retrieval-augmented generation

The RAG module plays a central role in improving the accuracy and reliability of natural language database querying. Rather than relying solely on the internal knowledge of a LLM, the RAG mechanism dynamically retrieves relevant contextual information from external knowledge sources, including database schemas, metadata descriptions, and entity relationships.

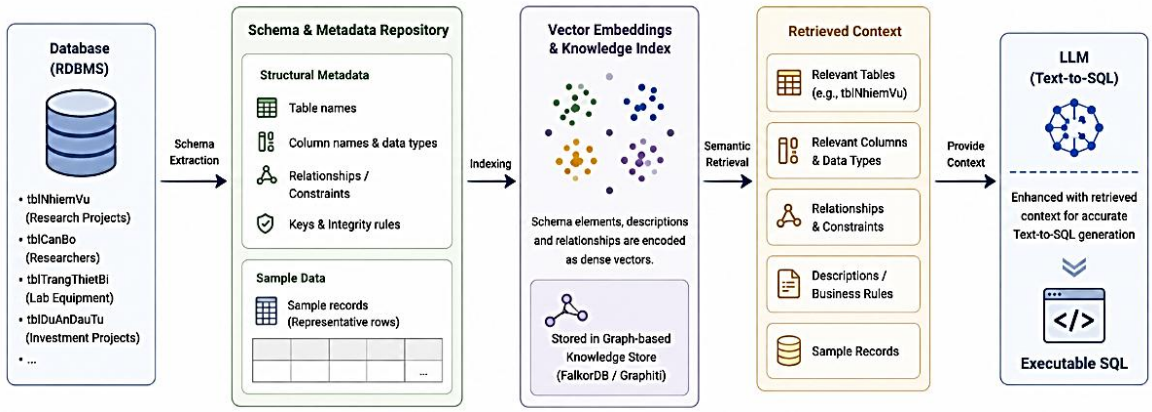

Figure 2. *Retrieval-Augmented Generation*

By incorporating domain-specific knowledge into the query generation process, RAG reduces semantic ambiguity, enhances schema comprehension, and improves the quality of Text-to-SQL translation. In the proposed framework, RAG serves as an intermediary layer that bridges user intentions and the underlying scientific research management database.

a) Schema extraction

The Schema Extraction module automatically collects and organizes database metadata to provide contextual knowledge for downstream SQL generation. During initialization, the system scans the relational database and extracts schema information, including table names, column names, data types, constraints, and representative sample records. For scientific research management databases, the extracted schema may include tables such as research projects (*tblNhiemVu*), researchers (*tblCanBo*), laboratory equipment (*tblTrangThietBi*), and investment projects (*tblDuAnDauTu*). This structured metadata repository provides explicit domain knowledge that facilitates accurate interpretation of user queries and reliable SQL generation.

b) Semantic schema linking

The Semantic schema linking module establishes semantic connections between user expressions and database entities. For instance, the terms research projects, researchers, laboratory equipment, and investment projects may correspond to the tables tblNhiemVu, tblCanBo, tblTrangThietBi, and tblDuAnDauTu, respectively. To support semantic matching, schema elements and metadata descriptions are transformed into dense vector embeddings and stored within a graph-based knowledge repository, such as FalkorDB or Graphiti. This enables semantic retrieval beyond keyword matching, allowing the system to identify relevant database entities even when users employ different terminology. Consequently, the framework improves schema selection accuracy and provides richer contextual information for subsequent Text-to-SQL generation.

2.3. LLM-based SQL generation

The LLM-based SQL generation module serves as the core component for transforming natural language queries into executable SQL statements. In the proposed framework, the LLM functions as a semantic translator that interprets user intent and generates database queries without requiring SQL expertise. The system supports both zero-shot and few-shot generation strategies. To enhance generation accuracy, prompt engineering is applied, incorporating task instructions, retrieved schema information, the user query, and a constrained output format. For example, for the query "List the top five research projects with the highest funding in 2024," the model uses information from tables such as *tblNhiemVu* and *tblKinhPhi* to generate an SQL statement that aggregates funding data, ranks projects by budget, and returns the top five results. This approach enables accurate and user-friendly database access through natural language interaction.

2.4. SQL execution, automatic error correction, and result presentation

Following SQL generation, the query is executed by the underlying Relational Database Management System (RDBMS), which validates the syntax, resolves schema relationships, and retrieves the requested information. To enhance reliability, the framework incorporates an automatic error-correction mechanism. When execution fails due to issues such as invalid table references, incorrect column names, or improper JOIN conditions, the generated SQL statement, database error message, and schema context are returned to the LLM for refinement. The model then produces a corrected SQL query, and the process is repeated automatically until successful execution is achieved or a predefined limit is reached. The retrieved results are presented as structured tabular data and subsequently analyzed by the LLM to generate concise natural language summaries. In addition, the system automatically selects and produces appropriate visualizations, including bar charts, pie charts, and line charts, based on the characteristics of the returned data. This integrated workflow improves query success rates, enhances result interpretability, and supports efficient exploration of scientific research management data through natural language interaction.

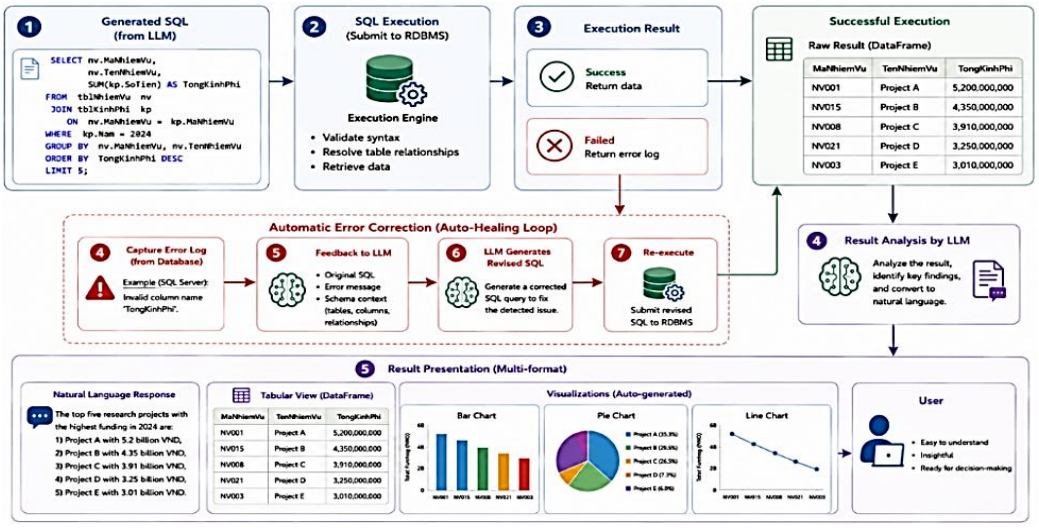

Figure 3. *SQL execution, automatic error correction, and result presentation.*

3. RESULTS AND DISCUSSION

3.1. Dataset construction

This study utilizes a Scientific Research Management Dataset developed by the Military Institute of Science and Technology. Based on this database, a domain-specific query dataset consisting of natural language–SQL pairs was constructed. These datasets provide a realistic benchmark for developing and evaluating Text-to-SQL and RAG techniques.

a) Scientific research management dataset

The Scientific Research Management Dataset is a relational database covering four main domains: researchers, research projects, investment projects, and laboratory equipment. It stores information on personnel profiles, project activities, funding, publications, infrastructure, and equipment management with 68 tables. The database contains numerous interconnected entities linked through foreign-key relationships, creating a realistic and complex environment for evaluating schema understanding, semantic retrieval, and Text-to-SQL generation in scientific research management applications.

b) Scientific research management query dataset

To evaluate the proposed framework, a benchmark dataset of 500 manually constructed natural language–SQL pairs was created by domain experts. The dataset is divided into three complexity levels: L1 (200 queries) containing single-table queries with simple filtering conditions, L2 (200 queries) requiring one or two JOIN operations, and L3 (100 queries) involving multiple JOINs and complex filtering conditions. Representative examples include retrieving researchers participating in funded projects during a specific period or identifying the highest-funded projects in a given year. This hierarchical design enables comprehensive assessment of schema linking, contextual retrieval, join prediction, and SQL generation performance across varying levels of query complexity.

Table 1. *Examples of natural query-SQL pairs in a query dataset.*

Level	Natural language question	SQL
L1	List all researchers who hold a PhD degree	SELECT * FROM tblCanBo WHERE HocViId = 3 ORDER BY HoTen;
L2	List all researchers who have the academic title of Professor.	SELECT cb.HoTen, hh.TenHocHam FROM tblCanBo cb LEFT JOIN tblHocHam hh ON hh.HocHamId = cb.HocHamId WHERE hh.TenHocHam = N'Giáo sư' ORDER BY cb.HoTen;
L3	List all female researchers with a PhD degree who work at the Institute of Information Technology.	SELECT cb.HoTen, hv.TenHocVi, gt.TenGioiTinh, dv.TenDonVi FROM tblCanBo cb LEFT JOIN tblHocVi hv ON hv.HocViId = cb.HocViId LEFT JOIN tblGioiTinh gt ON gt.GioiTinhId = cb.GioiTinhId LEFT JOIN tblDMDonVi dv ON dv.DMDonViId = cb.DMDonViId WHERE hv.TenHocVi = N'Tiến sĩ' AND gt.TenGioiTinh = N'Nữ' AND dv.TenDonVi = N'Viện Công nghệ Thông tin' ORDER BY cb.HoTen;

3.2. Experiment setup

To evaluate the effectiveness of the proposed framework, a complete Text-to-SQL system was implemented and tested using three large language models (LLMs): Qwen3-VL-32B-Instruct-FP8, Meta-Llama-3-70B-Instruct, and Gemma-2-9B. To clarify the hardware configuration and support reproducibility, these LLMs were deployed on a server equipped with NVIDIA A100 GPUs. A CPU-based client machine was then used to access the models via API to generate the SQL queries. Specifically, both the generated SQL and the corresponding ground-truth SQL were converted into vector representations using OpenAI's text-embedding-3-large model. Their semantic similarity was then measured using cosine similarity:

$$SimilarityScore(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|} \quad (1)$$

where $A \cdot B$ denotes the dot product and $\|\cdot\|$ represents the Euclidean norm. The similarity score

ranges from −1 to 1, with higher values indicating stronger semantic alignment. In general, scores above 0.9 represent highly similar query semantics, whereas scores below 0.7 indicate potentially incorrect SQL generation.

3.3. Results and discussion

Table 2. *Examples of queries and SQL generated by the proposed method.*

Natural language question	**SQL benchmark**	**SQL generated by proposed method**
List all researchers who hold a PhD degree	SELECT * FROM tblCanBo WHERE HocViId = 3 ORDER BY HoTen;	SELECT cb.* FROM tblCanBo cb INNER JOIN tblHocVi hv ON hv.HocViId = cb.HocViId WHERE hv.TenHocVi = N'Tiến sĩ' ORDER BY cb.HoTen;
Show all research projects whose implementation status is completed	SELECT NhiemVuId, TenNhiemVu, BatDau, KetThuc FROM tblNhiemVu WHERE TrangThaiThucHien = 5 ORDER BY KetThuc DESC;	SELECT NhiemVuId, TenNhiemVu, BatDau FROM tblNhiemVu WHERE TrangThaiThucHien = 5 ORDER BY BatDau DESC;
List all research projects belonging to the KC.07.18 Program.	SELECT nv.NhiemVuId, nv.TenNhiemVu, ct.TenChuongTrinhDeAn FROM tblNhiemVu nv LEFT JOIN tblChuongTrinhDeAn ct ON ct.ChuongTrinhDeAnId = nv.ChuongTrinhDeAnId WHERE ct.TenChuongTrinhDeAn = N' 'Chương trình KC.07.18' ORDER BY nv.NhiemVuId DESC;	SELECT nv.MaSo,nv.TongKinhPhi, ct.TenChuongTrinhDeAn FROM tblNhiemVu nv LEFT JOIN tblChuongTrinhDeAn ct ON ct.ChuongTrinhDeAnId = vv.ChuongTrinhDeAnId WHERE ct.TenChuongTrinhDeAn = N'Chương trình KC.07.18' ORDER BY nv.NhiemVuId DESC;
Find research projects at the ministry management level that are currently being implemented and belong to the field of Information Technology.	SELECT nv.NhiemVuId, nv.TenNhiemVu, cq.TenCapQuanLy, lv.TenLinhVucKH, tt.TenTinhTrang FROM tblNhiemVu nv LEFT JOIN tblCapQuanLy cq ON cq.CapQuanLyId = nv.CapQuanLyId LEFT JOIN tblLinhVucKH lv ON lv.LinhVucKHId = nv.LinhVucKHId LEFT JOIN tblTinhTrang tt ON tt.TinhTrangId = nv.TinhTrangId WHERE cq.TenCapQuanLy = N' Bộ Quốc phòng - Bộ, ngành khác ' AND lv.TenLinhVucKH = N'Công nghệ thông tin' AND nv.TrangThaiThucHien = 2 ORDER BY nv.NhiemVuId DESC;	SELECT nv.NhiemVuId, nv.TenNhiemVu, cq.TenCapQuanLy, lv.TenLinhVucKH, tt.TenTinhTrang FROM tblNhiemVu nv LEFT JOIN tblCapQuanLy cq ON cq.CapQuanLyId = nv.LoaiHinhId LEFT JOIN tblLinhVucKH lv ON lv.LinhVucKHId = nv.LinhVucKHId LEFT JOIN tblTinhTrang tt ON tt.TinhTrangId = nv.TinhTrangId WHERE cq.TenCapQL = N'Bộ Quốc Phòng - Bộ, ngành khác' AND lv.TenLinhVucKH = N'Công nghệ thông tin' AND nv.TrangThaiThucHien = 2 ORDER BY nv.NhiemVuId DESC;

The experimental results demonstrate the efficacy of the proposed Text-to-SQL framework in generating semantically accurate queries for the scientific management database. As shown in Table 2, the model successfully processes simple queries (L1, L2) —such as retrieving researchers with PhDs or completed projects—preserving the intended semantics despite variations in attribute selection and join structures. This underscores the limitations of exact string matching for evaluation. However, performance notably degrades in complex Level 3 (L3) queries that require multi-table reasoning and domain-specific knowledge. An in-depth failure analysis of these L3 queries reveals three primary sources of error. First, schema mapping errors frequently occur when the model selects incorrect table references or invalid attributes. Second, multi-table join errors arise from mismatched foreign keys during complex relational linkages. Finally, the model occasionally produces SQL syntax errors, rendering the generated queries unexecutable.

To quantitatively assess semantic correctness, cosine similarity between embeddings of generated and benchmark SQL queries was computed, where scores >0.9 denote high semantic alignment and <0.7 suggest potentially incorrect SQL generation. As shown in Table 3, Meta-Llama-3-70B-Instruct exhibited the highest overall robustness (0.83) and stability across complexity levels. However, trend analysis reveals that no model achieved the ideal >0.9 threshold. Furthermore, as query complexity increased to Level 3, the performance of Qwen3 and Gemma-2 dropped below the 0.7 critical threshold (0.65 and 0.67, respectively), indicating a high probability of generating invalid SQL structures in complex multi-table reasoning scenarios.

Table 3. *Similarity scores of generated SQL and standard SQL with different LLMs.*

LLM	Level 1	Level 2	Level 3	All
Qwen3-VL-32B-Instruct-FP8	0.8	0.69	0.65	0.76
Meta-Llama-3-70B-Instruct	**0.86**	**0.85**	**0.81**	**0.83**
Gemma-2-9B	0.72	0.68	0.67	0.71

4. CONCLUSIONS

This study presents an offline intelligent framework for scientific research management data exploitation by integrating LLM with RAG. The proposed system enables natural language queries to be automatically translated into SQL statements, allowing users without database expertise to access and analyze relational data efficiently. Deployment on local infrastructure ensures data security and operational independence, making the framework suitable for sensitive research environments. Experimental results demonstrate the effectiveness of the approach, with Meta-Llama-3-70B-Instruct achieving the highest semantic similarity score of 0.83, confirming the potential of open-source LLMs for domain-specific Text-to-SQL applications. Despite these promising results, performance declines as query complexity increases, particularly for Level 3 queries involving multiple joins and complex filtering conditions. Future work will focus on integrating auto-healing mechanisms through recursive feedback from database execution logs, enabling iterative SQL refinement. In addition, improvements to the RAG component, including enhanced schema linking and graph-based knowledge representations, will be investigated to reduce hallucinations and strengthen reasoning capabilities in complex real-world scenarios.

Acknowledgment: *This work was supported by the Institute-Level Research Project of the Military Institute of Science and Technology project, code: ĐTVCN.01.24/CNTT.*

REFERENCES

[1]. S. Adiyono and et al., "*AI and Digital Transformation Trends: A Systematic Review with Multi-Criteria Analysis*", Journal of Information Systems and Informatics, Vol. 8, No. 1, pp. 1011–1044, (2026).

[2]. T. Chowdhury, "*Cloud-Based Information Retrieval for Big Data: A Survey of Architectures and Scalability Challenge*", The Eastasouth Journal of Information System and Computer Science, Vol. 3, No. 3, pp. 312–322, (2026).

[3]. Z. Hong et al., "*Next-generation database interfaces: A survey of llm-based text-to-sql*", IEEE Transactions on Knowledge and Data Engineering, (2025).
[4]. Z. Luo, Z. Yang, Z. Xu, W. Yang and X. Du, "*Llm4sr: A survey on large language models for scientific research*", arXiv preprint arXiv:2501.04306, (2025).
[5]. M. Klesel and H. F. Wittmann, "*Retrieval-Augmented Generation (RAG)*", Business & Information Systems Engineering, Vol. 67, pp. 551–561, (2025). DOI: 10.1007/s12599-025-00945-3.
[6]. T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li et al., "*Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task*", Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3911–3921, (2018).
[7]. T. H. Nguyen, T. T. Bui and N. H. H. Luong, "*A study on deep learning for Vietnamese text classification*", Journal of Military Science and Technology, Vol. 95, pp. 85–94, (2024). DOI: 10.54939/1859-1043.j.mst.95.2024.85-94.
[8]. S. Gupta, R. Ranjan and S. N. Singh, "*A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions*", arXiv preprint arXiv:2410.12837, (2024).
[9]. P. P. Mishra, K. P. Yeole, R. Keshavamurthy, M. B. Surana and F. Sarayloo, "*A Systematic Framework for Enterprise Knowledge Retrieval: Leveraging LLM-Generated Metadata to Enhance RAG Systems*", arXiv preprint arXiv:2512.05411, (2025).
[10]. L. C. Chen, M. S. Pardeshi, Y. X. Liao and K. C. Pai, "*Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model*", Computer Standards & Interfaces, Vol. 94, p. 103995, (2025).
[11]. H. Thakkar and A. Manimaran, "*Comprehensive examination of instruction-based language models: A comparative analysis of mistral-7b and llama-2-7b*", 2023 International Conference on Emerging Research in Computational Science (ICERCS), IEEE, pp. 1–6, (2023).
[12]. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang et al., "*Training language models to follow instructions with human feedback*", Advances in Neural Information Processing Systems, Vol. 35, pp. 27730–27744, (2022).
[13]. T. H. Phan, M. P. Nguyen, H. H. Nguyen, T. T. T. Pham and T. L. Le, "*Tabular text embedding for Vietnamese text-based person search*", Journal of Military Science and Technology, Vol. 93, pp. 128–136, (2024). DOI: 10.54939/1859-1043.j.mst.93.2024.128-136.

TÓM TẮT

Nghiên cứu phát triển hệ thống thông minh ngoại tuyến phục vụ khai thác dữ liệu quản lý ngành khoa học công nghệ sử dụng mô hình ngôn ngữ lớn và kỹ thuật truy xuất tăng cường sinh (RAG)

Bài báo trình bày nghiên cứu phát triển hệ thống thông minh ngoại tuyến phục vụ khai thác dữ liệu quản lý khoa học và công nghệ trong môi trường mạng nội bộ có yêu cầu cao về bảo mật thông tin. Hệ thống được xây dựng trên cơ sở kết hợp các mô hình ngôn ngữ lớn (LLM) mã nguồn mở với kỹ thuật Retrieval-Augmented Generation (RAG), cho phép người dùng truy vấn, thống kê và tổng hợp dữ liệu bằng ngôn ngữ tự nhiên tiếng Việt mà không cần kiến thức về ngôn ngữ truy vấn cơ sở dữ liệu. Nghiên cứu tập trung vào ba nội dung chính: (1) xây dựng mô hình Text-to-SQL nhằm chuyển đổi câu hỏi tự nhiên thành câu lệnh SQL trên cơ sở dữ liệu chuyên ngành; (2) tích hợp cơ chế RAG để truy xuất lược đồ cơ sở dữ liệu, từ điển dữ liệu và các quy tắc nghiệp vụ, qua đó nâng cao khả năng hiểu ngữ cảnh và độ chính xác của truy vấn; và (3) triển khai các mô hình LLM hoạt động hoàn toàn trên hạ tầng nội bộ nhằm bảo đảm tính độc lập, giảm phụ thuộc vào dịch vụ đám mây và hạn chế nguy cơ rò rỉ dữ liệu. Ba mô hình Qwen 3, Meta-Llama và Gemma được đánh giá trên bộ dữ liệu truy vấn chuyên ngành. Kết quả thực nghiệm cho thấy hệ thống hoạt động ổn định trong môi trường ngoại tuyến, hỗ trợ hiệu quả truy vấn tiếng Việt tự nhiên, nâng cao khả năng khai thác dữ liệu và đáp ứng yêu cầu an toàn thông tin đối với cơ sở dữ liệu chuyên ngành.

Từ khoá: Text-to-SQL; Sinh tăng cường truy xuất (RAG); Mô hình ngôn ngữ lớn; Quản lý ngành khoa học công nghệ.