DOI: https://doi.org/10.15276/hait.09.2026.19
UDC 004.8:004.89:378.147

Context-aware educational chatbots via retrieval-augmented generation

The Vinh Tran[1)]
ORCID: https://orcid.org/0000-0002-4241-1065; vinhtt@ut.edu.vn. Scopus Author ID: 56835334700
Thi Khanh Tien Nguyen[1)]
ORCID: https://orcid.org/0000-0001-5379-7226; tienntk@ut.edu.vn. Scopus Author ID: 58735109800
Tran Nhat Linh Pham[1)]
ORCID: https://orcid.org/0009-0008-9245-3387; linhptn9746@ut.edu.vn. Scopus Author ID: 60219414200
[1)] Ho Chi Minh City University of Transport, 2, Vo Oanh Str. Ho Chi Minh City, 700000, Vietnam

ABSTRACT

Relevance: The rapid digital transformation of higher education has created growing demand for context-aware academic support. Existing educational chatbots suffer from poor grounding and high hallucination rates. **Purpose.** This study develops a context-aware educational chatbot framework that improves reliability and contextual relevance in bilingual academic environments. **Tasks:** The tasks include constructing a bilingual corpus, contrastively fine-tuning a multilingual retrieval model, and evaluating the framework against baselines. **Methods:** Course documents are converted into a structured Vietnamese–English question–answer corpus via semantic chunking, automated question–answer generation and paraphrase augmentation. A Transformer-based bi-encoder is contrastively fine-tuned and indexed in a Facebook AI Similarity Search vector database. Retrieved passages and conversation history are combined before response generation. **Scientific novelty:** The study integrates a data-centric pipeline with retrieval-augmented generation for bilingual education and introduces contrastive fine-tuning of a compact multilingual bi-encoder for retrieval tasks. **Practical significance:** The framework delivers a scalable, resource-efficient solution for intelligent tutoring with potential for institutional deployment. **Results:** Experimental evaluation demonstrates strong retrieval accuracy and contextual relevance. The system retrieves the most relevant content in nearly eighty-seven percent of cases and retrieves relevant information within the top results in more than ninety-five percent of evaluations. Compared with retrieval-free systems, the proposed framework improves factual consistency and reduces inaccurate responses by more than one-half while achieving stronger contextual relevance and semantic similarity than a standard baseline. **Conclusions:** Structured corpus construction and retrieval quality are key factors affecting chatbot performance. The proposed compact retrieval-augmented framework consistently improves factual grounding and contextual relevance in bilingual educational settings.

Keywords: Context-aware chatbots; retrieval-augmented generation; semantic embeddings; bilingual language processing; intelligent tutoring

For citation: **Tran T. V., Nguyen T. K. T., Pham T. N. L. "Context-aware educational chatbots via retrieval-augmented generation".** ***Herald of Advanced Information Technology.*** **2026; Vol. 9 No. 3: 289–306. DOI: https://doi.org/10.15276/hait.09.2026.19**

INTRODUCTION

The ongoing digital transformation of higher education has created a strong demand for scalable, personalised and always-available academic support tools. Intelligent conversational agents (chatbots) have emerged as a promising solution, providing instant responses, reducing administrative workload and improving access to learning resources. In e-learning environments, they increasingly function as virtual teaching assistants, guiding students through course materials and clarifying academic regulations [1]. Early chatbot systems relied on handcrafted rules and keyword matching, limiting them to narrow scenarios. The shift to machine learning and deep learning, particularly Transformer-based large language models (LLMs), has significantly improved language understanding and generation, enabling fluent and context-aware responses [2], [3].

limitations: difficulty maintaining multi-turn coherence, lack of access to institution-specific knowledge (e.g., syllabi, policies), and susceptibility to hallucinations – issues that directly impact reliability and student trust [4].

Retrieval-Augmented Generation (RAG) has been proposed to address these challenges by combining semantic retrieval with generative models [5]. In this framework, relevant documents are retrieved and incorporated into the input context before response generation, enabling factual grounding and updatable knowledge bases. Nevertheless, applying RAG to educational settings introduces additional challenges, including semi-structured and bilingual documents, complex tabular formats and context-dependent queries that require careful handling during retrieval.

To address these issues, this paper proposes a context-aware educational chatbot architecture that integrates deep natural language processing with retrieval-augmented generation under a data-centric

paradigm. The approach emphasizes dataset quality, semantic representation fidelity and retrieval effectiveness as key factors for system performance. The system is evaluated on a bilingual Vietnamese–English corpus constructed from 38 course syllabi in the Data Science program at the University of Transport Ho Chi Minh City, using both automatic metrics and human evaluation.

The main contributions are:

(1) a unified context-aware chatbot architecture combining DeepNLP and RAG;

(2) a structure-aware dataset construction pipeline for transforming academic documents into a bilingual QA corpus;

(3) fine-tuning a compact multilingual bi-encoder using contrastive learning and integrating it with a vector database and LLM; and

(4) a comprehensive evaluation against LLM-only and baseline RAG systems, including an ablation study.

LITERATURE REVIEW AND PROBLEM STATEMENT

The development of automatic dialogue systems is closely linked to advances in natural language processing and artificial intelligence. In education, chatbots have become key components of e-learning platforms and intelligent tutoring systems, supporting knowledge retrieval, task management and personalized learning [1], [6], [7], [8], [9], [10], [11], [12]. However, building reliable educational chatbots requires not only fluent responses but also context awareness and factual accuracy. A chronological review of chatbot architectures helps identify current limitations and motivates the proposed approach (Fig. 1a).

Rule-based chatbots. Rule-based systems represent the earliest generation, relying on handcrafted rules such as keyword matching and predefined scripts. A classic example is ELIZA [13], which uses pattern matching to simulate conversation. While simple and controllable, these systems lack learning capability, fail to capture deep semantics and scale poorly, making them unsuitable for complex educational settings.

Machine-learning-based chatbots. Statistical learning approaches, including Naive Bayes, SVM and decision trees, were introduced to automatically learn patterns from data, along with dialogue state tracking methods such as POMDP [14], [15]. These systems improve flexibility but still depend on handcrafted features and process utterances independently, limiting their ability to handle long-term context (Fig. 1b).

Deep learning-based chatbots.

eep learning models such as RNNs, LSTMs and sequence-to-sequence architectures significantly enhance language representation and enable end-to-end dialogue generation [16], [17], [18], [19]. However, they struggle with long contexts and lack mechanisms for incorporating external knowledge.

Transformer architecture and large language models

The Transformer architecture [2] enables parallel computation and effective modelling of long-range dependencies through self-attention. Pre-trained models such as BERT [20] and RoBERTa [21] have become standard tools for language understanding, while Vietnamese-specific models such as PhoBERT [22] further improve performance on monolingual tasks. Sentence-level embedding models like Sentence-BERT [23] support semantic similarity tasks, and efficient variants such as MiniLM [24] and MPNet [25] reduce computational cost while maintaining competitive performance. Transformer-based models have also been applied to diverse NLP tasks [26], [27].

More recently, instruction-tuned large language models (LLMs) have demonstrated strong capabilities in natural language generation [3], [28]. Compared with encoder-only architectures, models such as Qwen 2.5-3B-Instruct [29], [30] provide superior performance in instruction following and multilingual generation, making them particularly suitable for bilingual Vietnamese–English scenarios. However, despite these advances, LLMs used in isolation still suffer from hallucination and lack reliable grounding, which limits their applicability in knowledge-intensive domains such as education [31].

Dense retrieval and vector-based methods

Dense retrieval techniques project queries and documents into a shared vector space, enabling efficient semantic search. Methods such as DPR [32], dual-encoders [33] and ColBERT [34], along with benchmarks like BEIR [35], demonstrate strong performance. Contrastive learning further improves embedding quality [36], [37].

Retrieval-augmented generation. Retrieval-Augmented Generation integrates retrieval with text generation by incorporating relevant documents into the input context [5]. Approaches such as REALM [38] and Fusion-in-Decoder [39] show that this design improves factual accuracy and reduces hallucination. Fig. 1c illustrates the standard RAG architecture.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

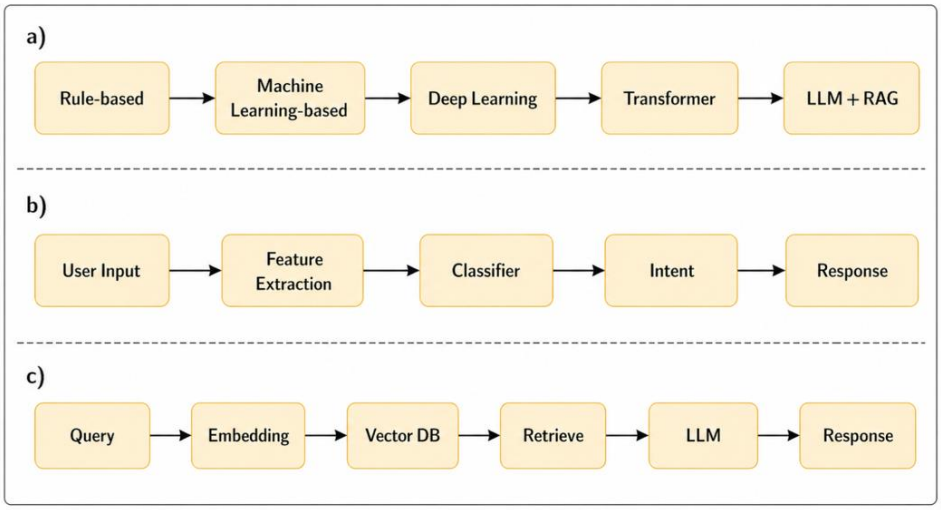

Fig. 1. **Combined overview of chatbot evolution and architectures: a – Evolution of chatbot systems from rule-based to LLMs; b – Pipeline of a machine learning based chatbot; c – Canonical Retrieval Augmented Generation architecture**

***Source*: compiled by the authors**

Comparative analysis and research gap

Building upon the evolution of chatbot paradigms discussed above, Table 1 provides a comparative summary of the main approaches in terms of context awareness, semantic representation, retrieval capability, integration with LLMs and suitability for educational applications. This comparison highlights a clear progression from rigid rule-based systems to more flexible learning-based and deep learning approaches, culminating in modern LLM- and RAG-based architectures.

Table 1. **Comparison of chatbot approaches**

Method	Context	Retrieval	LLM	Limitation
Rule-based	Weak	No	No	Rigid
ML-based	Partial	No	No	Manual features
Deep learning	Moder-ate	No	No	Weak long context
LLM-only	Strong	No	Yes	Hallucin-ation
RAG-based	Strong	Yes	Yes	Retrieval depend-ency
Proposed	Strong	Optimi-zed	Yes	Mitigated

***Source:* compiled by the authors**

Despite these advances, several limitations remain. Pure LLM-based systems, while fluent, lack reliable grounding mechanisms and are prone to hallucinations. Retrieval-Augmented Generation improves factual accuracy, yet existing implementations are typically not tailored for educational settings, where data are often semi-structured, domain-specific and bilingual. In addition, most approaches treat semantic representation, retrieval and generation as loosely coupled components, without jointly optimizing them for end-to-end performance.

Motivated by these limitations, this work proposes a context-aware educational chatbot architecture that integrates deep natural language processing and retrieval-augmented generation within a unified, data-centric framework. The proposed approach explicitly addresses the challenges of bilingual academic data, long conversational context and knowledge grounding, thereby bridging the gap between general-purpose conversational models and domain-specific educational requirements.

RESEARCH AIM AND OBJECTIVES

The aim of this research is to design, implement and empirically evaluate a context-aware educational chatbot that integrates deep natural language processing with retrieval-augmented generation to improve response accuracy, factual grounding and contextual consistency in bilingual academic support scenarios.

To achieve this aim, the following objectives are defined:

1) analyse the current landscape of educational chatbots and identify key limitations of rule-based, machine learning, deep learning and LLM-only approaches;

2) design a unified context-aware architecture that combines DeepNLP and RAG, tailored for bilingual Vietnamese–English academic content;

3) develop a data-centric pipeline for constructing a high-quality bilingual question–answer corpus from semi-structured course syllabi;

4) fine-tune a compact multilingual Transformer bi-encoder using contrastive learning and integrate it with a vector database and a large language model;

5) conduct a comprehensive evaluation using retrieval, generation and human-centred metrics, including an ablation study to quantify the contribution of each component.

The scientific novelty of this work is concentrated in two contributions: (i) the contrastive fine-tuning of a compact multilingual Transformer bi-encoder specifically for bilingual Vietnamese–English educational retrieval, demonstrating that domain-adapted compact models can match or exceed larger pre-trained alternatives; and (ii) the empirical validation of retrieval quality and embedding fidelity as the primary performance drivers in educational retrieval-augmented generation systems, establishing a reproducible bilingual evaluation framework. The data-centric pipeline for structured corpus construction (objective 3) constitutes an engineering contribution that operationalizes these scientific claims – it is a designed implementation artefact rather than a principally new algorithmic method.

MATERIALS AND METHODS

System overview

The proposed system is built upon a standard Retrieval-Augmented Generation (RAG) pipeline: User Query → Embedding → Vector Database → Context Retrieval → Large Language Model (LLM) → Response Generation. The framework is organised into three core components: dataset construction, semantic representation learning, and retrieval-augmented response generation. Unlike conventional RAG approaches that directly apply pre-trained models to raw documents, this work adopts a data-centric paradigm in which dataset quality and embedding fidelity are the primary determinants of system performance.

The architecture is structured as a two-phase framework (Fig. 2), separating offline knowledge preparation from online inference.

Phase 1: Knowledge Base Construction and Semantic Indexing (offline).

This phase focuses on preparing and structuring domain knowledge prior to deployment. Its objectives are to construct a structured academic dataset, learn high-quality semantic representations, and build an efficient vector index for fast retrieval. The pipeline is defined as: Raw Documents → Preprocessing → Semantic Chunking → QA Generation → Embedding → Vector Index (FAISS).

Phase 2: Context-Aware Retrieval-Augmented Inference (online).

This phase processes user queries in real time, aiming to interpret user intent, retrieve relevant knowledge, and generate accurate, context-aware responses. The inference pipeline is: User Query → Query Embedding → Retrieval → Context Construction → LLM → Response.

The two-phase design improves computational efficiency while clearly decoupling knowledge representation learning from inference. This separation enhances scalability, stability and reliability, making the system well suited for academic chatbot applications where the knowledge base is updated periodically.

Dataset construction

The dataset is built from thirty-eight course syllabi of the Data Science program at the University of Transport Ho Chi Minh City, provided in .docx format. Each document contains academic information such as course objectives, course learning outcomes (CLOs), teaching contents and assessment methods. The input data present several characteristics that make direct processing difficult: they are semi-structured, contain multi-level tables with complex relationships, include checkboxes and special symbols, and mix Vietnamese and English text. The corpus is intentionally scoped to a single academic program to maintain a controlled and coherent knowledge domain for evaluation purposes.

The dataset construction pipeline consists of five main stages:

Stage 1 – Document Extraction (.docx → JSON). Data are extracted directly from the XML structure of .docx files to preserve all textual content, tables, and embedded symbols. The intermediate output is converted into a structured JSON format.

Stage 2 – Semantic Segmentation. Each document is partitioned into self-contained knowledge units based on section titles, logical structure, and content type. Each segment is annotated with metadata, including course name, section type, and information category.

Stage 3 – Question-Answer (QA) Generation. Large language models (e.g., Gemini, ChatGPT) are employed to generate questions corresponding to the answer content within the constructed knowledge base, thereby restructuring the data into a QA format aligned with user query behavior.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

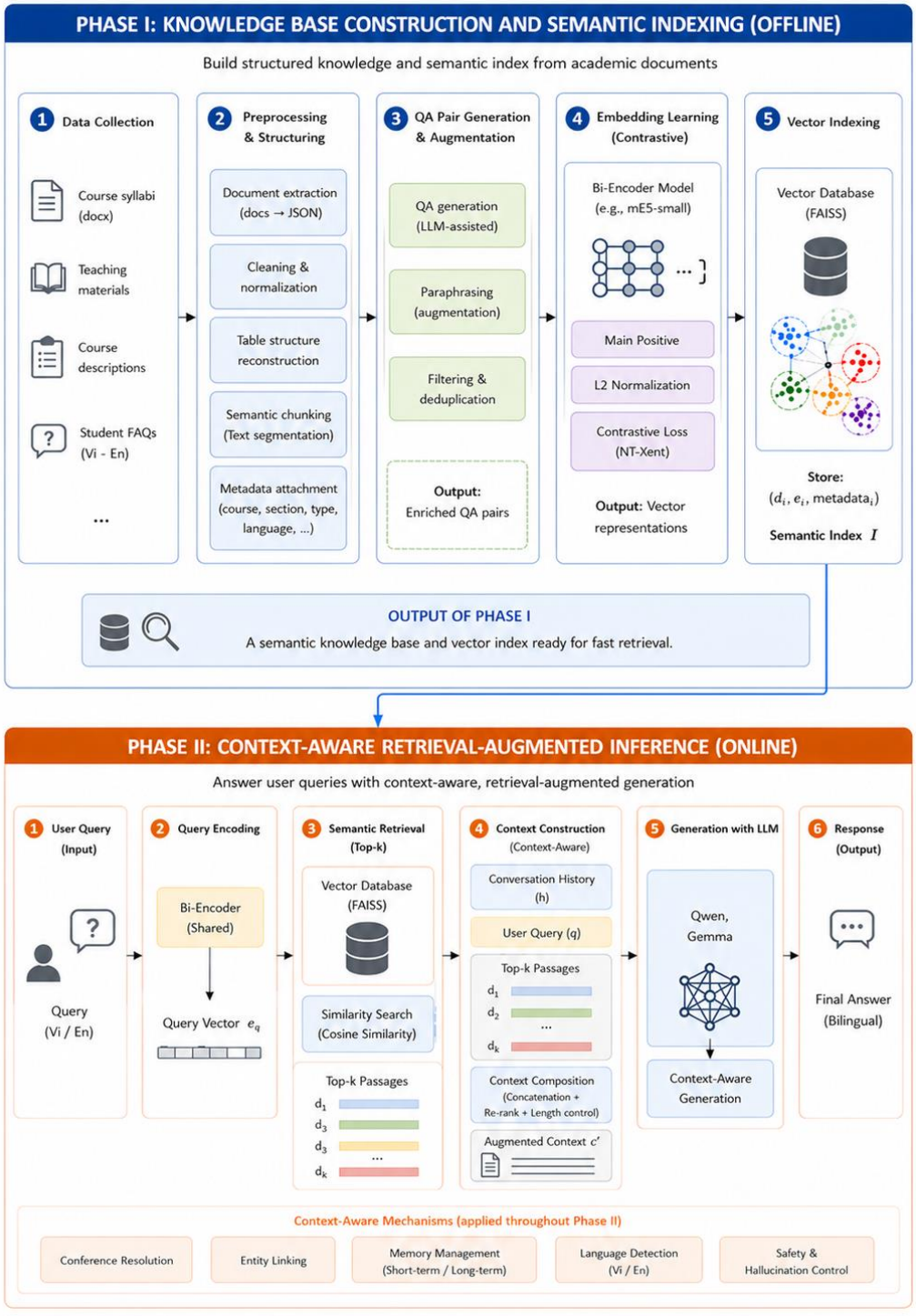

Fig. 2. **Two-phase architecture of the unified context-aware educational chatbot using deep NLP and RAG for bilingual (Vietnamese-English) academic content**
Source: **compiled by the authors**

Stage 4 – Data Augmentation. A prompting-based approach is applied using Gemini-2.5-Flash-Lite to generate semantically equivalent paraphrases, expanding the query space and increasing data diversity.

Stage 5 – Preprocessing and Normalization. The dataset is refined through the removal of invalid artifacts, text normalization, restructuring of tabular content, and token-length control. Finally, the processed data are standardized into JSON format for downstream training and retrieval tasks.

To improve dataset quality and semantic consistency, generated bilingual question–answer pairs, segmented knowledge units, and metadata annotations were manually reviewed during corpus construction. This verification process aimed to

remove duplicated, inconsistent, or semantically ambiguous samples before downstream training and retrieval tasks.

Semantic embedding learning (Phase 1 – Retrieval)

To represent queries and documents in a shared semantic space, the system adopts a Transformer-based bi-encoder as the core component of Phase 1 in the retrieval pipeline. As illustrated in Fig. 3, the architecture consists of two parallel encoding branches that process the query $q$and each candidate document d_i, respectively. Both branches share identical parameters (shared weights), ensuring that query and document representations are projected into a unified embedding space.

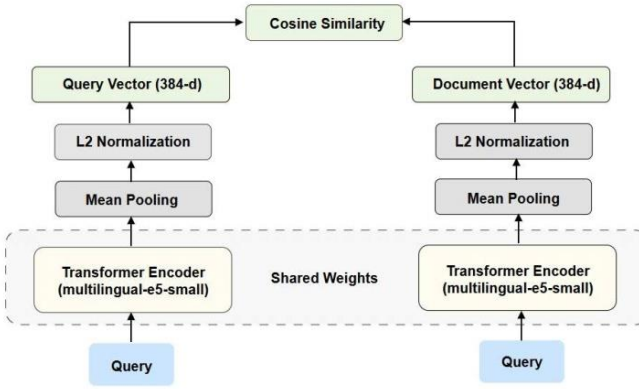

Fig. 3. **Bi-Encoder architecture based on the multilingual-e5 model**

Source: **adapted by the author based on Wang et al. [33] and Reimers & Gurevych [19]**

Each input sequence is first tokenized and fed into a pretrained Transformer encoder (*multilingual-e5-small*), which produces contextualized token embeddings. These token-level representations are then aggregated using a mean-pooling operation to obtain fixed-dimensional sentence embeddings, followed by L2 normalization to stabilize similarity computation.

Formally, the embedding vectors are defined as:

$$e_q = f_{embed}(q)\,,$$
$$e_{d_i} = f_{embed}(d_i) \in R^d\,, \quad (1)$$

where f_{embed} denotes the shared encoder function and d is the embedding dimension.

The multilingual-e5-small model was selected as the base encoder based on three practical considerations: multilingual coverage, parameter efficiency, and inference efficiency in bilingual retrieval settings. Instruction-tuned encoders such as Instructor require task-specific prompts during inference and provide less suitable support for Vietnamese–English semantic retrieval. Although GTE variants achieve strong English benchmark performance, their multilingual coverage and Vietnamese representation remain comparatively limited for the targeted educational corpus. Larger E5 variants also introduce substantially higher computational cost without providing stable retrieval improvements after domain-specific adaptation. In contrast, multilingual-e5-small offers broad multilingual pre-training across more than 100 languages, including Vietnamese and English, while maintaining a compact architecture and efficient inference characteristics. Collectively, these properties position multilingual-e5-small as an effective and computationally efficient backbone for the proposed bilingual retrieval-augmented framework and its subsequent domain-specific fine-tuning process.

Relevance between the query and candidate documents is computed using cosine similarity:

$$sim(e_q, e_i) = (e_q \cdot e_i) / (\|e_q\| \, \|e_i\|)\,. \quad (2)$$

The encoder is fine-tuned using a contrastive learning objective. Within each mini-batch B containing a positive document d^+, the NT-Xent loss is defined as:

$$L = -log\left(\frac{exp\left(\frac{sim(q,d^+)}{\tau}\right)}{\Sigma_{\{d\in B\}} exp\left(\frac{sim(q,d)}{\tau}\right)}\right)\,, \quad (3)$$

where τ denotes the temperature hyperparameter. In-batch negatives are employed to avoid the need for explicit hard negative mining.

This training objective encourages semantically related query–document pairs to be mapped closer in the embedding space while pushing unrelated pairs further apart.

The model is optimized using the AdamW optimizer with a learning rate of 2.10^{-5}, a batch size of 16, a maximum sequence length of 256 tokens, and 50 training epochs.

Vector database and retrieval

All document embeddings are stored in a FAISS index, which supports efficient approximate nearest-neighbor search in high-dimensional spaces.

For a given query embedding e_q, the system retrieves the top-k documents with the highest cosine similarity:

$$R_k = Top - k\{\, d_i \,|\, sim(e_q, e_i)\,\}\,, \quad i = 1,..,N\,. \quad (4)$$

The parameter k controls the trade-off between retrieval coverage and prompt-length constraints. A sensitivity analysis was conducted to determine an appropriate range for the educational use case. At k=1, only the single most similar passage is

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

retrieved, minimizing prompt length and latency but risking incomplete coverage when a query spans multiple course topics. Increasing k to 3 consistently captures the primary answer passage alongside at least one supporting passage while keeping the total prompt well within the 2048-token context limit. At k=5, coverage reaches its practical ceiling for single-topic educational queries: additional passages beyond this point tend to duplicate already-retrieved content or introduce tangentially related material that dilutes generation focus and marginally increases latency. Values in the range three to five were therefore found to maximise faithfulness and context relevance while preserving low inference latency, and this range is adopted as the operational setting throughout all experiments.

Context augmentation

The retrieved documents are combined with the user query to build an augmented context.

When conversation history h is available, the system forms a context-aware prompt:

$$c' = concat(h, q, d_1, d_2, \ldots, d_k), \quad (5)$$

Three heuristics are applied to stabilize the prompt: relevance-based ordering of documents, removal of noisy or redundant passages, and enforcement of a maximum token budget.

These steps are crucial in education, where the clarity and structure of supporting information directly influence answer quality.

In the current experimental setup, conversation history h is treated as an optional component and all evaluation scenarios are conducted in a single-turn setting in which h is empty. The effect of multi-turn dialogue on retrieval quality and response faithfulness is therefore not directly measured in this study and is identified as a direction for future investigation.

Context-aware Retrieval-augmented Response Generation Algorithm

Input: User query q, dataset D, conversation history h, top-k parameter k

Output: Generated response y

Offline Phase: Knowledge Indexing

- Step1: Preprocess dataset D to obtain cleaned documents $\{d_i\}$,
- Step2: Compute document embeddings:
$$e_i = f_{\text{embed}}(d_i),\ \forall d_i \in D,$$
- Step 3: = Store and index $\{e_i\}$in a vector database (e.g., FAISS)

Online Phase: Context-Aware Inference

- Step 4: Compute query embedding:
$$e_q = f_{\text{embed}}(q)$$
- Step 5: Retrieve top-k relevant documents:
$$R_k = \text{Top-}k\{d_i \mid \text{sim}(e_q, e_i)\}$$
- Step 6: Construct augmented context:
$$c' = \text{concat}(h, q, d_1, \ldots, d_k)$$
- Step 7: Generate response using LLM:
$$y = f_{\text{LLM}}(c')$$
- Step 8: Return y

Steps 1-3 are executed once during system initialization, while Steps 4-8 are performed for each incoming query.

RESEARCH RESULTS

Experimental setup

The experiments were conducted on a GPU-enabled server equipped with an Intel Xeon E5-2680 v4 CPU (2.40 GHz)**,** 62 GB RAM**,** and an NVIDIA GeForce RTX 5070 Ti (16 GB) GPU. The proposed system integrates the multilingual-e5-small encoder for semantic embedding, a FAISS-based vector database for efficient retrieval, and a large language model for response generation.

The multilingual-e5-small encoder produces 384-dimensional embeddings while the number of retrieved documents is configured in the range of $k = 3$ to 5. The maximum input context length for the LLM is limited to 2048 tokens to ensure computational efficiency, providing a balance between retrieval accuracy, inference latency, and memory usage for real-time deployment.

After data augmentation, the final corpus comprises 28,101 question–answer pairs, which are split into training **(**22,481 samples**)** and validation **(**5,620 samples) sets using an 80/20 ratio, as reported in Table 2. The training hyperparameters used during fine-tuning are summarized in Table 3.

Table 2. **Dataset partition used for embedding model fine-tuning**

Dataset	Samples
Training	22 481
Validation	5 620
Total	28 101

Source: compiled by the authors

Dataset quality evaluation

The quality of the constructed corpus is evaluated along four complementary dimensions: scale, coverage, token-length distribution, and query–answer length correlation.

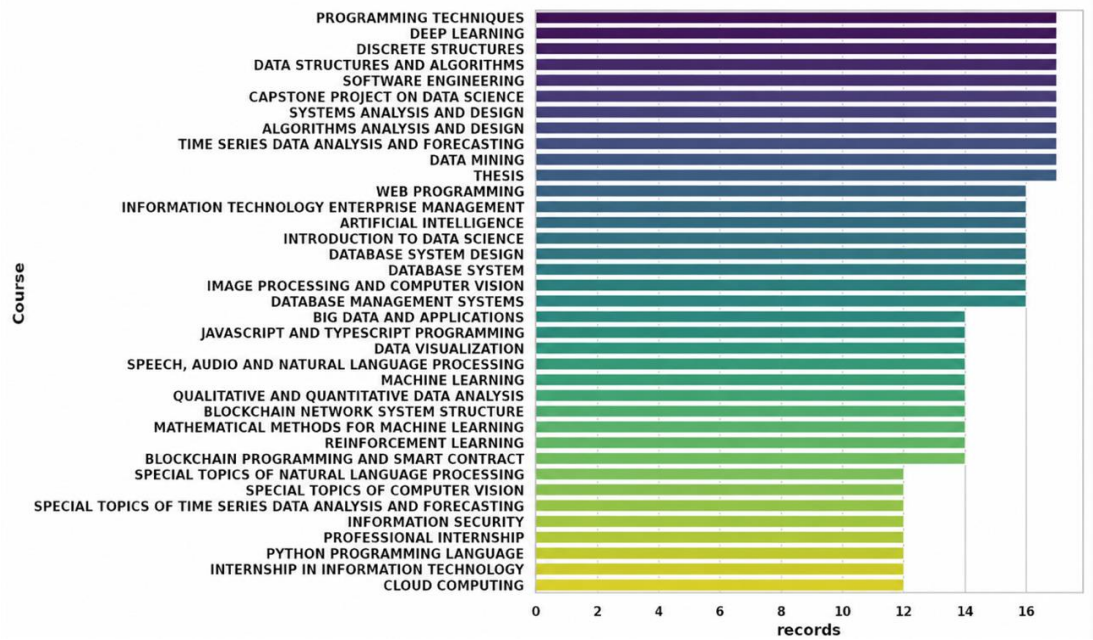

Fig. 4. **Distribution of question-answer samples across courses**

Source: **compiled by the authors**

Table 3. **Training hyperparameters for the embedding model**

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	$2x10^{-5}$
Batch size	16
Epochs	50
Max sequence length	256
Temperature τ	0.05

Source: **compiled by the authors**

First, the augmentation pipeline expands the dataset by more than fifty times (from 551 to 28,101 samples) compared to the original syllabi. This substantial increase enhances the density of the embedding space, thereby improving nearest-neighbor search and reducing retrieval variance.

Second, the dataset achieves an approximately balanced distribution across the 38 courses (Fig. 4), which helps mitigate retrieval bias and ensures stable performance across the knowledge domain.

Third, the token-length distribution (Fig. 5) shows that queries are generally short (mostly under 100 tokens), consistent with real user behavior, while answers span a broader range (typically 100–300 tokens), with some exceeding 512 tokens. To handle long responses, a length-aware chunking strategy is applied to prevent information loss during embedding.

Finally, the analysis in Fig. 6 reveals no clear linear correlation between query and answer lengths, indicating that answer length is primarily driven by content complexity rather than query size.

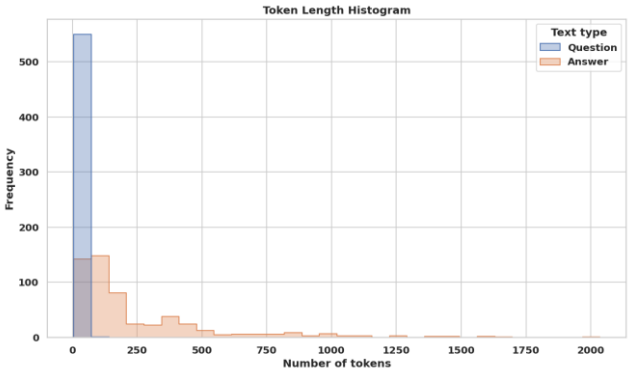

Fig. 5. **Token-length distribution of questions and answers**

Source: **compiled by the authors**

Overall, these results demonstrate that the dataset is sufficiently large, well-balanced, and statistically consistent, making it suitable for training embedding models and supporting robust semantic retrieval.

Embedding-model training and retrieval performance

Fig. 7 presents the training and validation loss curves of the multilingual-e5-small encoder during fine-tuning. The model converges rapidly, the loss decreases smoothly and the gap between training and validation losses remains small, which indicates good generalization and no evidence of severe overfitting.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

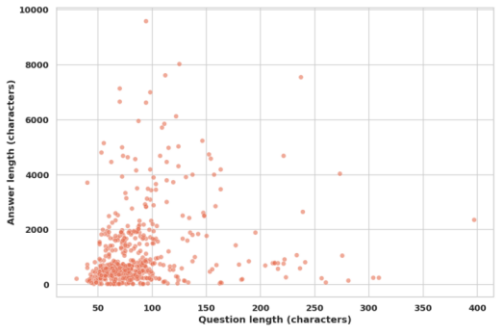

Fig. 6. **Relationship between question length and answer length**

Source: **compiled by the authors**

Table 4 shows the trade-off between retrieval latency and parameter count. The fine-tuned multilingual-e5-small achieves a response time of approximately 12.11 ms per query, substantially faster than mE5-large (≈22.71 ms) while maintaining superior accuracy. Such low latency is well below the perceptual threshold of human users and allows the system to handle concurrent requests at reasonable computational cost.

Table 4. **Latency and parameter efficiency of embedding models**

Model	Inference Time (ms)	Parameters (millions)
mE5-large	22.71	559.9
PhoBERT-large	22.54	369.2
mE5-base	13.41	278
MiniLM	12.42	117.7
PhoBERT-base	12.27	135
mE5-small	**12.11**	**117.7**
MPNet	11.88	278

Source: **compiled by the authors**

Table 5 reports Recall@K for the proposed model in comparison with the MiniLM, MPNet, PhoBERT and larger mE5 baselines. The proposed multilingual-e5-small achieves Recall@1 of 0.868, Recall@5 of 0.938 and Recall@10 of 0.955, outperforming both lighter and heavier baselines and demonstrating that model size alone is not the determining factor; alignment between training objective, multilingual coverage and the domain of the data is more important.

Table 5. **Recall@K of embedding models on the educational retrieval task**

Model	Recall@1	Recall@5	Recall@10
MPNet (Finetuned)	0.783	0.931	0.95
MPNet (Original)	0.174	0.424	0.559
MiniLM (Finetuned)	0.791	0.904	0.934
MiniLM (Original)	0.149	0.363	0.477
PhoBERT-base (Finetuned)	0.689	0.851	0.903
PhoBERT-base (Original)	0.056	0.132	0.182
PhoBERT-large (Finetuned)	0.765	0.892	0.92
PhoBERT-large (Original)	0.068	0.17	0.23
mE5-base (Finetuned)	0.705	0.844	0.902
mE5-base (Original)	0.487	0.796	0.874
mE5-large (Finetuned)	0.718	0.889	0.926
mE5-large (Original)	0.659	0.881	0.922
mE5-small (Finetuned)	**0.868**	**0.938**	**0.955**
mE5-small (Original)	0.472	0.784	0.87

Source: **compiled by the authors**

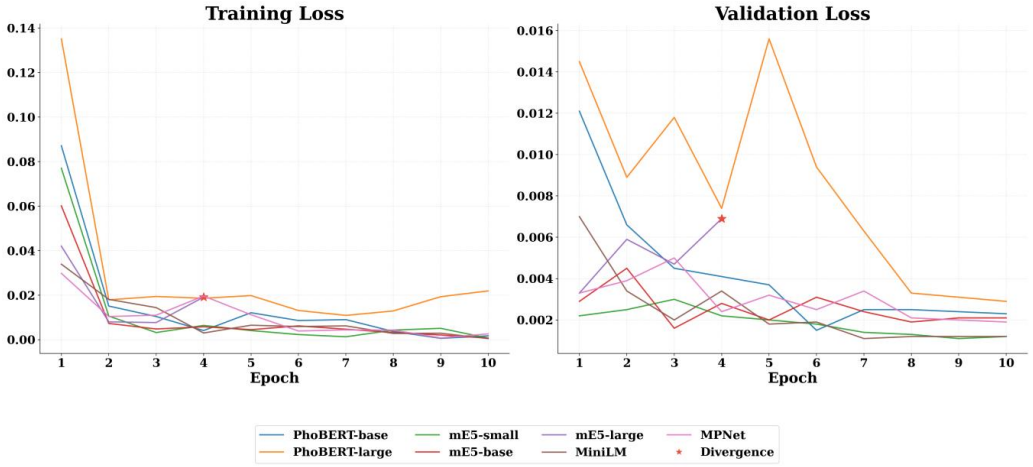

Fig. 7. **Training and validation loss of the fine-tuned bi-encoder**

Source: **compiled by the authors**

End-to-end performance comparison

Table 6 presents the end-to-end evaluation of four progressively enhanced chatbot configurations: (i) an LLM-only configuration without retrieval grounding, (ii) a Standard RAG (RAG baseline) configuration with top-5 retrieval, and (iii) the proposed Data-centric RAG framework integrating structured corpus preprocessing, semantic augmentation, contrastive fine-tuning of the multilingual-e5-small encoder, and optimized retrieval for bilingual educational content. The evaluation is conducted on a held-out set of 300 query–answer pairs from the validation split.

Response-level evaluation metrics, including faithfulness, contextual relevance, and hallucination detection, were computed automatically using the DeepEval framework [40] rather than manual annotator scoring. DeepEval employs LLM-based evaluation strategies to estimate semantic consistency between generated responses and retrieved context, contextual alignment with user queries, and the presence of unsupported or fabricated content in generated outputs. Although automated evaluation improves scalability and reproducibility, the evaluation process may still inherit limitations and potential biases from the underlying evaluator models.

The generation model is Qwen2.5-3B-Instruct, while an LLM-as-a-judge protocol with Qwen2.5-7B-Instruct is employed to assess responses along three reference-free dimensions: faithfulness to retrieved context, hallucination rate, and context relevance. In addition, lexical overlap is measured using ROUGE-1, ROUGE-2, and ROUGE-L F1, semantic similarity using BERTScore F1, and efficiency via average inference latency (seconds per query). In Table 6, Faithfulness denotes faithfulness (higher is better), the hallucination rate (lower is better), and Context Relevance. the context relevance (higher is better). R-1, R-2, and R-L correspond to ROUGE scores, BERTSc. to BERTScore F1, and Time (s) to average inference latency.

The proposed system (data-centric RAG) achieves the best performance across all quality metrics. Compared with the LLM-only baseline, faithfulness increases substantially from 0.492 to 0.752 (+0.260), while hallucination is reduced from 0.960 to 0.430 (−55.2%). ROUGE-L improves from 0.259 to 0.472 and BERTScore from 0.888 to 0.924, indicating that retrieval effectively grounds responses in authoritative course content.

Compared with the standard RAG, the most notable gain is observed in context relevance, which increases from 0.353 to 0.668 (+0.315). This confirms that the data-centric pipeline and fine-tuned bi-encoder retrieve semantically more relevant passages than a generic encoder.

The slight increase in latency (20.63s vs. 19.92s) reflects additional context processing but remains within the same practical range.

These results, further illustrated in Fig. 8, demonstrate that the proposed approach consistently improves both factual reliability and contextual alignment without compromising computational feasibility.

Expert validation and qualitative analysis

To further validate the effectiveness of the proposed system, retrieved passages and generated responses were examined through expert review. A domain expert (a faculty member with direct knowledge of the course syllabi corpus) independently assessed whether the retrieved content was relevant, accurate, and sufficient to ground the generated response for a sample of queries. Two representative cases from this review are presented below to illustrate system behaviour compared with baseline approaches.

Case Study 1: Hallucination in prerequisite queries.

Query: “What is the general information about the Cloud Computing course?”

Table 6. **End-to-end performance comparison of chatbot configurations 300 samples — inference LLM: Qwen2.5-3B-Instruct – judge LLM: Qwen2.5-7B-Instruct**

Method	Faithful-ness	Hallucin-ation	Context Relevance	R-1	R-2	R-L	BERT score	Time (s)
LLM-only	0.492	0.960	0.628	0.416	0.191	0.259	0.888	17.59
Standard RAG (k=5)	0.742	0.450	0.353	0.544	0.336	0.373	0.899	19.92
Data-centric RAG (proposed)	0.752	0.430	0.668	0.561	0.479	0.472	0.924	20.63

Source: **compiled by the authors**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

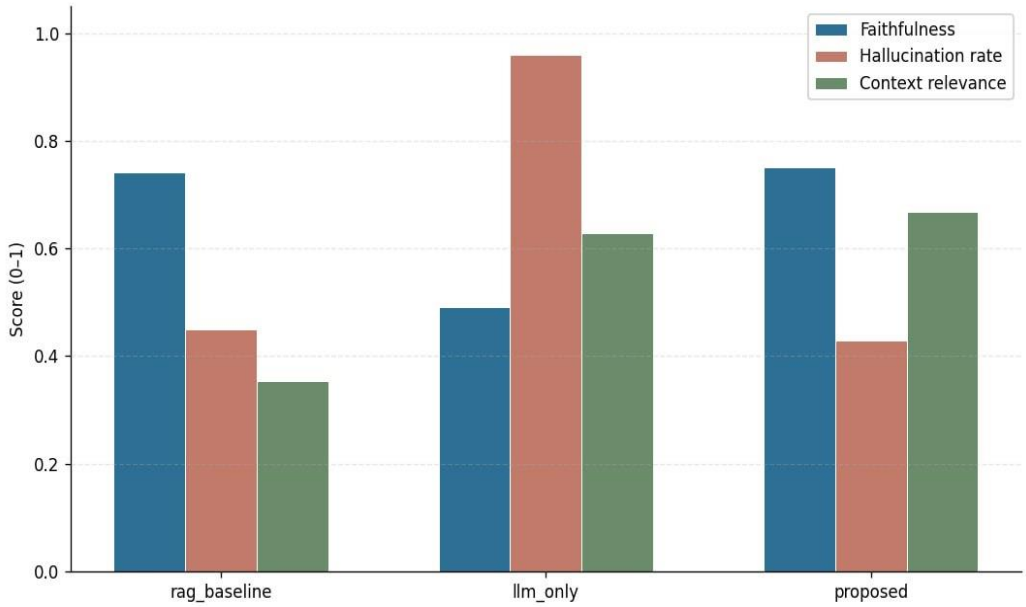

Fig. 8. **Faithfulness, hallucination rate and context relevance of the three chatbot configurations**
Source: **compiled by the authors**

The ground truth, obtained from the official syllabus, provides structured information including course name, credit allocation, total hours, prerequisite requirements, and course classification.

The LLM-only model generates a fluent but factually incorrect response by introducing generic and non-existent content, such as cloud technologies, tools, and programming languages, which are not present in the original syllabus. This behavior reflects a typical hallucination issue, where the model relies on prior knowledge rather than grounded evidence.

In contrast, the proposed system retrieves the relevant syllabus segment and produces a response that is fully aligned with the source content, accurately preserving key attributes such as credit distribution, prerequisite course, and course structure. This example demonstrates the system's ability to generate factually grounded responses and explains the significant reduction in hallucination rate observed in the quantitative evaluation (from 0.960 to 0.430).

Case Study 2: Context relevance in concept explanation.

Query: "What are the assessment methods for the Data Science Practical Project course?"

The ground truth specifies a structured evaluation scheme, including both process assessment (50 %) and final assessment (50%), along with detailed criteria and corresponding learning outcomes.

The LLM-only model again produces a generic and loosely related answer, describing common evaluation practices (e.g., essays, multiple-choice tests, and interviews) without grounding in the actual course specification.

The baseline RAG system partially retrieves relevant information but introduces inconsistencies in weight distribution and includes irrelevant details (e.g., exam formats and grading levels), indicating limitations in retrieval precision and context selection.

By contrast, the proposed system generates a response that is more precise and contextually relevant, correctly reflecting the assessment structure and associated evaluation criteria. This improvement is attributed to the fine-tuned retriever, which selects more accurate and task-relevant context from the knowledge base. As a result, the system achieves a significantly higher context relevance score (0.668 compared to 0.353 for the baseline RAG model).

Overall, these case studies confirm that the performance improvements of the proposed system are not only reflected in quantitative metrics but also in qualitative behavior. Specifically, the system demonstrates a stronger ability to produce factually grounded responses and maintain contextual relevance. The analysis highlights the critical role of high-quality retrieval in mitigating hallucination and enhancing answer accuracy in educational chatbot systems. As illustrated in Fig. 9 , the query "What are the assessment methods for the Data Science Practical Project course?" is first embedded, then used to retrieve relevant syllabus segments, and finally passed to the LLM to generate a grounded response.

Curl

```
curl -X 'POST' \
  'http://localhost:8000/query/stream' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "query": "What are the assessment methods for the Data Science Practical Project course?",
  "top_k": 5,
  "threshold": 0.5
}'
```

Request URL

```
http://localhost:8000/query/stream
```

Server response

Code	Details
200	**Response body**

```
| **General Information**           |
|-----------------------------------|
| Course Code: SDE007               |
| Credits: 4                        |
| **Course Learning Objectives**    |
|-----------------------------------|
| Learning Outcomes:                |
| - CLO1: Apply data science skills and tools to solve real-world problems |
| - CLO2: Develop and manage data science projects                          |
| - CLO3: Build and evaluate data-driven predictive models                 |
| **Detailed Content**              |
|-----------------------------------|
- Process assessment accounts for 50% of the total grade, including:
    - 5% attendance
    - 45% periodic group report presentations
- Final assessment accounts for 50% of the total grade, including:
    - Project report presentation
- Both process and final assessments apply evaluation criteria A1.2 and A5.4,
  to ensure learning outcomes CLO1, CLO2, and CLO3 are met.
```

Fig. 9. **Query processing pipeline in the proposed system**

Source: **compiled by the authors**

Ablation study

The ablation study uses k=1 and k=5 as boundary conditions within the recommended range: k=1 isolates the minimum retrieval benefit – single document grounding; while k=5 represents the upper limit of the range, quantifying the maximum gain from multi-document retrieval before context-window and latency constraints become binding.

Table 7 reports an ablation study that isolates the contribution of progressively enhanced retrieval and data-centric optimization components within the proposed framework. The faithfulness score increases monotonically across the four configurations: from 0.492 for the LLM-only configuration without retrieval grounding, to 0.708 for Standard RAG with top-1 retrieval, to 0.742 for Standard RAG with top-5 retrieval, and finally to 0.752 for the proposed Data-centric RAG framework. Conversely, the hallucination rate decreases substantially from 0.960 to 0.430 across the same progression. The largest improvement is observed when retrieval is first introduced; confirming that external knowledge grounding is the primary factor in reducing hallucinations and improving factual reliability. Increasing the retrieval depth from k = 1 to k = 5 further improves contextual coverage, while the addition of the data-centric preprocessing and domain-adaptive fine-tuning pipeline provides additional gains in both faithfulness and context relevance.

Table 7. **Ablation study of the proposed pipeline**

Configuration	Faithful-ness	Hallucin-ation	Context Relevance
LLM-only	0.492	0.960	0.628
Standard RAG (k=1)	0.708	0.510	0.602
Standard RAG (k=5)	0.742	0.450	0.353
Data-centric RAG	**0.752**	**0.430**	**0.668**

Source: **compiled by the authors**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

DISCUSSION OF RESULTS

The experimental results demonstrate that the proposed context-aware chatbot consistently outperforms both the LLM-only and Standard RAG baselines across all evaluation metrics, validating the five research objectives set out in Section 3. The most substantial gain is observed in context relevance, which increases from 0.353 (Standard RAG) to 0.668 (+0.315) – the largest absolute improvement across all metrics – indicating markedly stronger alignment between retrieved content and generated responses. Faithfulness improves from 0.492 (LLM-only) to 0.752 (+0.260), while the hallucination rate decreases from 0.960 to 0.430. These gains are particularly consequential in educational settings, where factual inaccuracy or unsupported claims can directly mislead students relying on the system for academic guidance.

Three key factors explain these improvements. First, the data-centric pipeline (Objective 3) transforms semi-structured Vietnamese – English course syllabi into a semantically coherent question–answer corpus of 28,101 pairs – an expansion of more than 50-fold relative to the original 551 source segments. This close alignment between training data distribution and real user query patterns contributes directly to generation quality: ROUGE-L increases from 0.259 (LLM-only) and 0.373 (Standard RAG) to 0.472, while BERTScore F1 rises from 0.888 and 0.899 to 0.924, reflecting higher lexical and semantic fidelity with respect to reference answers.

Second, the contrastive fine-tuning of the multilingual bi-encoder (Objective 4) produces a well-structured embedding space tailored to the bilingual educational domain. The fine-tuned multilingual-e5-small achieves Recall@1 of 0.868 and Recall@10 of 0.955, outperforming all baselines including substantially larger models such as mE5-large (Recall@1: 0.718; 559.9M parameters) and mE5-base (0.705; 278M parameters). This finding indicates that domain-specific contrastive fine-tuning is a more effective strategy than scaling model size for retrieval in specialized educational corpora, and that compact models can achieve superior performance when training data is well-aligned with the target domain. The fine-tuned encoder also maintains a low inference latency of 12.11 ms per query, making it suitable for deployment in interactive academic environments.

Third, combining retrieval with a generative model (Objective 2) provides factual grounding that is absent in standalone LLM systems. The LLM-only baseline, despite producing fluent responses, exhibits a hallucination rate of 0.960, reflecting its reliance on parametric knowledge without access to institution-specific content. The proposed system reduces this rate to 0.430 by conditioning response generation on verified syllabus passages retrieved at query time. The additional computation introduced by retrieval and context augmentation results in a moderate increase in end-to-end latency (20.63 s vs. 17.59 s for the LLM-only baseline), a trade-off that is acceptable in non-real-time educational query settings where response accuracy takes precedence over speed.

The ablation study (Table 7) provides further insight into the relative contribution of each component. Removing retrieval entirely reduces faithfulness from 0.752 to 0.492 (−0.260), confirming that retrieval is the dominant factor in system reliability. A counterintuitive but important observation is that vanilla RAG with k=5 yields lower context relevance (0.353) than k=1 (0.602). This is explicable by retrieval quality: when the encoder is not fine-tuned for the target domain, increasing the number of retrieved passages introduces semantically marginal or irrelevant content that dilutes the context and degrades query–response alignment. By contrast, the proposed system – equipped with a domain-adapted encoder – achieves context relevance of 0.668 even at k=5, demonstrating that retrieval quality, rather than retrieval quantity, is the critical determinant of contextual alignment. Incorporating the data-centric preprocessing pipeline yields further incremental gains in faithfulness and hallucination reduction, highlighting the complementary nature of encoder quality and corpus quality.

Compared with prior work in educational chatbots [1], [6] and retrieval-augmented generation [5], [38], [39], the present findings reinforce two observations that are consistent across the literature. First, generic RAG systems based on off-the-shelf encoders tend to suffer from limited context relevance when applied to domain-specific corpora, as reflected in the baseline score of 0.353. Second, domain adaptation – through both retrieval fine-tuning and targeted corpus curation – is essential for reliable performance in knowledge-intensive settings. The proposed approach addresses this gap by jointly optimizing retrieval and dataset quality within a unified bilingual framework, yielding consistent improvements across all retrieval, generation, and response-level metrics.

Despite these contributions, several limitations should be acknowledged. Regarding dataset scope, the training corpus is derived exclusively from 38

course syllabi of a single Data Science programmed at one institution. This homogeneity restricts generalizability: performance may vary across curricula from different disciplines, educational levels, or institutional terminologies. Furthermore, the corpus does not encompass diverse document types commonly found in broader educational environments – such as lecture slides, textbooks, assessment rubrics, or online discussion forums – which may contain knowledge structures substantially different from those of formal course syllabi.

Regarding evaluation scope, the test set of 300 QA pairs, while sufficient for comparative analysis, constrains statistical confidence in cross-metric comparisons. Human evaluation involved a single domain expert reviewing a targeted sample of retrieved passages and generated responses; a more rigorous methodology would require multiple annotators with formal inter-rater agreement measures such as Cohen's kappa. All experiments are additionally conducted in a single-turn dialogue setting, leaving the cumulative effect of multi-turn conversation history on retrieval accuracy, faithfulness, and context relevance unevaluated.

Regarding system scope, the generative component (Qwen2.5-3B-Instruct) is treated as a fixed black box; its internal behavior under adversarial queries, out-of-distribution inputs, or code-switched Vietnamese – English prompts is not fully characterized. The evaluation additionally relies on automated LLM-as-a-judge scoring via Qwen2.5-7B-Instruct, which may inherit systematic biases from the evaluator model and may not capture all nuances that expert human annotators would identify. Future work will address these limitations by validating the pipeline across multiple institutions and document formats, introducing dedicated multi-turn evaluation protocols, expanding the expert evaluation panel, and characterizing the generative component's behavior under adversarial conditions.

Overall, the results confirm that the performance of retrieval-augmented educational chatbots is primarily determined by retrieval quality, and that domain-specific fine-tuning of the retrieval encoder provides substantially greater benefit than increasing model size. The proposed framework – integrating a contrastive fine-tuned bilingual bi-encoder, a curated domain-specific QA corpus, and a retrieval-augmented generative pipeline – addresses all five stated research objectives and demonstrates that compact, domain-adapted models can achieve strong and consistent performance in specialized educational settings. These findings have direct practical implications for academic institutions seeking cost-effective, scalable, and factually reliable conversational support systems.

Ethical considerations

The deployment of educational chatbots raises several ethical considerations that warrant explicit attention. First, responsibility for incorrect or misleading responses is a key concern: while the proposed system grounds responses in verified syllabus documents, residual hallucinations remain possible, and institutions should establish clear policies positioning the chatbot as a supplementary tool rather than an authoritative source. Instructors and institutions, rather than the system itself, should retain ultimate responsibility for the accuracy of academic guidance provided to students. Second, transparency is essential: students interacting with the system should be explicitly informed that responses are generated by an artificial intelligence model and may require independent verification. Third, the corpus is derived from institutional documents that may reflect existing biases in curriculum design, such as underrepresentation of certain student populations or learning styles; institutions should audit system outputs to ensure the chatbot does not inadvertently reinforce inequitable educational practices. Fourth, query logs may contain identifiable student information and must be handled in accordance with applicable data protection regulations. These considerations do not diminish the utility of the proposed system but highlight the importance of responsible deployment practices in educational contexts.

CONCLUSIONS

This study proposed a context-aware educational chatbot framework that addresses the limitations of both standalone large language model systems and generic retrieval-augmented generation architectures in bilingual academic environments. By pursuing the five research objectives defined in Section 3, the work delivered: (i) a systematic analysis of chatbot paradigm evolution and its limitations; (ii) a two-phase unified architecture integrating deep natural language processing with retrieval-augmented generation; (iii) a structure-aware pipeline constructing a 28,101-pair bilingual Vietnamese-English question–answer corpus from 38 course syllabi; (iv) a contrastive fine-tuned multilingual bi-encoder integrated with a FAISS vector database for semantic retrieval; and (v) a comprehensive evaluation across retrieval, generation, and response-level metrics. The resulting system achieved Recall@1 of 0.868, surpassing all

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

baselines including models with substantially higher parameter counts, while improving faithfulness from 0.492 to 0.752, reducing the hallucination rate from 0.960 to 0.430, and raising context relevance from 0.353 to 0.668 – the largest absolute gain across all evaluated dimensions.

The study advances the field through two principal scientific contributions. First, the contrastive fine-tuning of the compact multilingual-e5-small encoder (117.7M parameters) for bilingual Vietnamese-English educational retrieval demonstrates that domain-specific training substantially outweighs the benefit of model scale: the fine-tuned compact model achieves Recall@1 of 0.868, surpassing both mE5-large (0.718; 559.9M parameters) and mE5-base (0.705; 278M parameters) while maintaining an inference latency of 12.11 ms per query. Second, the structure-aware dataset construction pipeline – which transforms semi-structured academic syllabi into a semantically coherent question-answer corpus through document extraction, semantic segmentation, LLM-assisted generation, and data augmentation – provides a replicable methodology for building domain-specific retrieval corpora from institutional documents, applicable beyond the Vietnamese educational context.

Together, these contributions demonstrate that high retrieval quality, achieved through domain adaptation rather than model scaling, is the primary driver of factual reliability in educational chatbot systems. The framework establishes that compact, resource-efficient architectures – combining a fine-tuned bi-encoder, a curated bilingual corpus, and a small generative model – can deliver strong and consistent performance in specialised academic settings without requiring large computational infrastructure. These findings carry direct practical implications for higher education institutions in resource-constrained environments seeking deployable, scalable, and factually reliable conversational support systems.

The present study is subject to constraints inherent to its experimental scope: the corpus is limited to a single academic programme at one institution, and all evaluations are conducted in single-turn interaction scenarios. Future work will address these limitations through three targeted directions: (i) extending the corpus and evaluation to cross-institutional and multi-disciplinary educational settings to assess generalisability; (ii) developing dedicated multi-turn evaluation protocols that incorporate persistent conversational memory; and (iii) investigating hybrid retrieval strategies combining dense and sparse signals to resolve the context relevance degradation observed when increasing the number of retrieved passages with weaker encoders. Validating the framework's pedagogical impact through real-world student interaction studies remains a longer-term objective.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the University of Transport Ho Chi Minh City for providing the syllabus materials used in this research.

REFERENCES

1. Okonkwo, C. W. & Ade-Ibijola, A. “Chatbots applications in education: A systematic review”. *Computers and Education: Artificial Intelligence*. 2021; 2: 100033, https://www.scopus.com/pages/publications/85120809017.DOI: https://doi.org/10.1016/j.caeai.2021.100033.

2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. 2017; 30: 5998–6008. DOI: https://doi.org/10.48550/arXiv.1706.03762.

3. Brown, T., Mann, B., Ryder, N., et al. “Language models are few-shot learners”. In *Advances in Neural Information Processing Systems*. 2020; 33: 1877–1901. DOI: https://doi.org/10.48550/arXiv.2005.14165.

4. Kasneci, E., Sessler, K., Küchemann, et al. “ChatGPT for good? On opportunities and challenges of large language models for education”. *Learning and Individual Differences*. 2023; 103: 102274, https://www.scopus.com/pages/publications/85150364293. DOI: https://doi.org/10.1016/j.lindif.2023.102274.

5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. & Kiela, D. “Retrieval-augmented generation for knowledge-intensive NLP tasks”. In *Advances in Neural Information Processing Systems*. 2020; 33: 9459–9474. DOI: https://doi.org/10.48550/arXiv.2005.11401.

6. Winkler, R. & Söllner, M. “Unleashing the potential of chatbots in education: A state-of-the-art analysis”. *Academy of Management Proceedings*. 2018; 1: 15903. DOI: https://doi.org/10.5465/AMBPP.2018.15903abstract.

7. Adamopoulou, E. & Moussiades, L. “Chatbots: History, technology, and applications”. *Machine Learning with Applications*. 2020; 2: 100006. DOI: https://doi.org/10.1016/j.mlwa.2020.100006.

8. Korchenko, O. G.; Tereikovskyi, I. A.; Korystin, O. Y.; Tereikovska, L. O.; Tereikovskyi, O. I. “A method for detecting financial phishing in instant messengers using an ensemble of dialogical intelligent assistants based on large language models”. *Herald of Advanced Information Technology*. 2026; 9 (1): 71–84, https://www.scopus.com/pages/publications/105031963942.
DOI: https://doi.org/10.15276/hait.09.2026.06.

9. Sheremet, O. I.; Sadovoi, O. V.; Sheremet, K. S.; Sokhina, Y. V. “Effective documentation practices for enhancing user interaction through GPT-powered conversational interfaces”. *Applied Aspects of Information Technology*. 2024; 7 (2): 135–150. DOI: https://doi.org/10.15276/aait.07.2024.10.

10. Kolesnikov, O. Y., Biloshchytskyi, A. O. & Faizullin, A. R. “Elaboration of theoretical foundations for the creation of the educational environment of the educational institution”. *Applied Aspects of Information Technology*. 2020; 3 (2): 32–43. DOI: https://doi.org/10.15276/aait.02.2020.2.

11. Makogon, H., Lavrut, T., Chornyi, M., Matuzko, B. & Załoga, W. “Information technology of the acquiring professional competencies process in the e-learning management system environment by construction an educational trajectory”. *Advanced Information Systems*. 2024; 8 (4): 103–117, https://www.scopus.com/pages/publications/85209191548. DOI: https://doi.org/10.20998/2522-9052.2024.4.13.

12. Antoshchuk, S. G. & Breskina, A. A. “Human action analysis models in Artificial Intelligence based proctoring systems and dataset for them”. *Applied Aspects of Information Technology*. 2023; 6 (2): 190–200. DOI: https://doi.org/10.15276/aait.06.2023.14.

13. Weizenbaum, J. “ELIZA – A computer program for the study of natural language communication between man and machine”. *Communications of the ACM*. 1966; 9 (1): 36–45. DOI: https://doi.org/10.1145/365153.365168.

14. Jurafsky, D. & Martin, J. H. “Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models”. 2026. – Available from: https://web.stanford.edu/~jurafsky/slp3. – [Accessed: Jan. 2026].

15. Young, S., Gašić, M., Thomson, B. & Williams, J. D. “POMDP-based statistical spoken dialogue systems: A review”. *Proceedings of the IEEE*. 2013; 101 (5): 1160–1179. DOI: https://doi.org/10.1109/JPROC.2012.2225812.

16. Hochreiter, S. & Schmidhuber, J. “Long short-term memory”. *Neural Computation*. 1997; 9 (8): 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735.

17. Sutskever, I., Vinyals, O. & Le, Q. V. “Sequence to sequence learning with neural networks”. In *Advances in Neural Information Processing Systems*. 2014; 27: 3104–3112. DOI: https://doi.org/10.48550/arXiv.1409.3215.

18. Lashyn, Y., Trofymchuk, O., Zabolotnyi, S., Voitko, O. & Seabra, E. “Sentiment analysis of texts using recurrent neural networks of the transformer architecture”. *Advanced Information Systems*. 2025; 9 (3): 91–101, https://www.scopus.com/pages/publications/105012831513. DOI: https://doi.org/10.20998/2522-9052.2025.3.11.

19. Bocharova, M. Y. & Malakhov, E. V. “ResJobFit – End-to-End artificial neural networks based technology for job-resume matching”. *Applied Aspects of Information Technology*. 2024; 7 (4): 378–391. DOI: https://doi.org/10.15276/aait.07.2024.27.

20. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of NAACL-HLT*. 2019. p. 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423.

21. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. “RoBERTa: A robustly optimized BERT pretraining approach”. *arXiv*. 2019. DOI: https://doi.org/10.48550/arXiv.1907.11692.

22. Nguyen, D. Q. & Nguyen, A. T. “PhoBERT: Pre-trained language models for Vietnamese”. In *Findings of EMNLP*. 2020. p. 1037–1042. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.92.

23. Reimers, N. & Gurevych, I. “Sentence-BERT: Sentence embeddings using Siamese BERT-networks”. In *Proceedings of EMNLP-IJCNLP*. 2019. p. 3982–3992. DOI: https://doi.org/10.18653/v1/D19-1410.

24. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N. & Zhou, M. “MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers”. In: *Advances in Neural Information Processing Systems*. 2020; 33: 5776–5788. DOI: https://doi.org/10.48550/arXiv.2002.10957.

25. Song, K., Tan, X., Qin, T., Lu, J. & Liu, T.-Y. “MPNet: Masked and permuted pre-training for language understanding”. In: *Advances in Neural Information Processing Systems*. 2020; 33: 16857–16867. DOI: https://doi.org/10.48550/arXiv.2004.09297.

26. Gavrylenko, S., Poltoratskyi, V. & Nechyporenko, A. “Intrusion detection model based on improved transformer”. *Advanced Information Systems*. 2024; 8 (1): 94–99, https://www.scopus.com/pages/publications/85186197574. DOI: https://doi.org/10.20998/2522-9052.2024.1.12.

27. Podorozhniak, A., Liubchenko, N., Oliinyk, V. & Roh, V. “Research application of the spam filtering and spammer detection algorithms on social media and messengers”. *Advanced Information Systems*. 2023; 7 (3): 60–66, https://www.scopus.com/pages/publications/85176957483. DOI: https://doi.org/10.20998/2522-9052.2023.3.09.

28. “OpenAI. GPT-4 technical report”. *arXiv*. 2023. DOI: https://doi.org/10.48550/arXiv.2303.08774.

29. Bai, J., Bai, S., Yang, Y., et al. “Qwen technical report”. *arXiv*. 2023. DOI: https://doi.org/10.48550/arXiv.2309.16609.

30. “Alibaba Cloud Team. Qwen2 technical report”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2407.10671.

31. Syromiatnikov, M. V. & Ruvinskaya, V. M. “A system internals modeling and annotation language for large language model-driven software engineering”. *Applied Aspects of Information Technology*. 2026; 9 (1): 103–121. DOI: https://doi.org/10.15276/aait.09.2026.08.

32. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. & Yih, W. “Dense passage retrieval for open-domain question answering”. In *Proceedings of EMNLP*. 2020. p. 6769–6781, https://www.scopus.com/pages/publications/85118435873. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.550.

33. Ni, J., Qu, C., Lu, J., Dai, Z., Hernandez Abrego, G., Ma, J., Zhao, V. Y., Luan, Y., Hall, K. B., Chang, M.-W. & Yang, Y. “Large dual encoders are generalizable retrievers”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2112.07899.

34. Khattab, O. & Zaharia, M. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT”. In *Proceedings of ACM SIGIR*. 2020: 39–48. DOI: https://doi.org/10.1145/3397271.3401075.

35. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. & Gurevych, I. “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models”. In *NeurIPS Datasets and Benchmarks Track*. 2021. DOI: https://doi.org/10.48550/arXiv.2104.08663.

36. Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J. & Overwijk, A. “Approximate nearest neighbor negative contrastive learning for dense text retrieval”. In *International Conference on Learning Representations*. 2021. DOI: https://doi.org/10.48550/arXiv.2007.00808.

37. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R. & Wei, F. “Text embeddings by weakly-supervised contrastive pre-training”. *arXiv*. 2022. DOI: https://doi.org/10.48550/arXiv.2212.03533.

38. Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M.-W. “REALM: Retrieval-augmented language model pre-training”. In: *Proceedings of the 37th International Conference on Machine Learning*. 2020. p. 3929–3938. DOI: https://doi.org/10.48550/arXiv.2002.08909.

39. Izacard, G. & Grave, E. “Leveraging passage retrieval with generative models for open-domain question answering”. In *Proceedings of EACL*. 2021. p. 874–880. DOI: https://doi.org/10.18653/v1/2021.eacl-main.74.

40. “Confident AI. DeepEval: The open-source LLM evaluation framework”. *GitHub repository*. 2023. – Available from: https://github.com/confident-ai/deepeval. – [Accessed: Jan. 2026].

Conflicts of Interest: The authors declare that they have no conflicts of interest regarding this study, including financial, personal, authorship, or other interests that could influence the research and the results presented in this article

Received 07.04.2026
Received after revision 12.06.2026
Accepted 19.06.2026

DOI: https://doi.org/10.15276/hait.09.2026.19
УДК 004.8:004.89:378.147

Контекстно-залежні освітні чат-боти за допомогою генерації з доповненим пошуком даних

Чан Тхе Вінь[1)]
ORCID: https://orcid.org/0000-0002-4241-1065; vinhtt@ut.edu.vn. Scopus Author ID: 56835334700
Нгуєн Тхі Кхань Тієн[1)]
ORCID: https://orcid.org/0000-0001-5379-7226; tienntk@ut.edu.vn. Scopus Author ID: 58735109800
Фам Чан Нят Лінь[1)]
ORCID: https://orcid.org/0009-0008-9245-3387; linhptn9746@ut.edu.vn
[1)] Університет транспорту міста Хошимін, вул. Во Оань, 2. м. Хошимін, 700000, В'єтнам

АНОТАЦІЯ

Актуальність: Швидка цифрова трансформація вищої освіти створила зростаючий попит на контекстно-залежну академічну підтримку. Існуючі освітні чат-боти характеризуються недостатньою прив'язкою до предметної області та високим рівнем галюцинацій. **Мета.** Дослідження спрямоване на розробку контекстно-залежного освітнього чат-бота, який підвищує надійність відповідей і контекстуальну релевантність у двомовному академічному середовищі. **Завдання:** Завдання включають побудову двомовного корпусу, контрастне налаштування багатомовної моделі пошуку та оцінювання платформи порівняно з базовими системами. **Методи:** Документи курсу перетворюються на структурований в'єтнамсько-англійський корпус запитань і відповідей за допомогою семантичного сегментування, автоматизованої генерації запитань і відповідей та перефразування. Бі-енкодер на основі Transformer контрастно налаштовується та індексується у векторній базі даних Facebook AI Similarity Search. Отримані фрагменти та історія діалогу об'єднуються перед генерацією відповідей. **Наукова новизна:** Дослідження поєднує конвеєр обробки даних із генерацією відповідей, доповненою пошуком, для двомовної освіти та впроваджує контрастне налаштування компактного багатомовного бі-енкодера для задач пошуку. **Практичне значення:** Платформа пропонує масштабоване та ресурсоефективне рішення для інтелектуального навчання з потенціалом інституційного впровадження. **Результати:** Експериментальна оцінка демонструє високу точність пошуку та контекстуальну релевантність. Система знаходить найбільш релевантний контент майже у вісімдесяти семи відсотках випадків і виявляє релевантну інформацію серед найвищих результатів більш ніж у дев'яноста п'яти відсотках оцінювань. Порівняно із системами без пошуку, запропонована структура покращує фактичну узгодженість і зменшує кількість неточних відповідей більш ніж наполовину, одночасно забезпечуючи вищу контекстуальну релевантність і семантичну подібність порівняно зі стандартною базовою системою. **Висновки:** Структурована побудова корпусу та якість пошуку є ключовими факторами, що впливають на ефективність чат-бота. Запропонована компактна структура з доповненим пошуком послідовно покращує фактичну достовірність і контекстуальну релевантність у двомовному освітньому середовищі.

Ключові слова: контекстно-орієнтовані чат-боти; генерація з доповненим пошуком; семантичні ембединги; двомовна обробка мови; інтелектуальне навчання

ABOUT THE AUTHORS

The Vinh Tran - Doctor of Philosophy, Program Director of Data Science. Ho Chi Minh City University of Transport, 2, Vo Oanh Str. Ho Chi Minh City, 700000, Vietnam
ORCID: https://orcid.org/0000-0002-4241-1065; vinhtt@ut.edu.vn. Scopus Author ID: 56835334700
Research field: Data Science, Cybersecurity, Blockchain Technologies, Deep Learning and Systems Architecture

Чан Тхе Вінь - доктор філософії, директор Програми з науки про дані, Університет транспорту міста Хошимін, , вул. Во Оань, 2. Хошимін, 700000, В'єтнам

Thi Khanh Tien Nguyen - Doctor of Philosophy, Lecturer. Ho Chi Minh City University of Transport, 2, Vo Oanh Str., Ho Chi Minh City, 700000, Vietnam
ORCID: https://orcid.org/0000-0001-5379-7226; tienntk@ut.edu.vn. Scopus Author ID: 58735109800
Research field: Data science, Deep Learning, Computer Vision, Natural Language processing

Нгуєн Тхі Кхань Тієн - доктор філософії, старший викладач. Університет транспорту міста Хошимін, вул. Во Оань, 2. Хошимін, 700000, В'єтнам

Tran Nhat Linh Pham - Student of Data Science major, Institute of Information Technology & Electrical and Electronics Engineering. Ho Chi Minh City University of Transport, 2, Vo Oanh Str. Ho Chi Minh City, 700000, Vietnam
ORCID: https://orcid.org/0009-0008-9245-3387; linhptn9746@ut.edu.vn. Scopus Author ID: 60219414200
***Research field*:** Data science, Deep Learning, Computer Vision, Natural Language processing

Фам Чан Нят Лінь - студент спеціальності «Наука про дані», Інститут інформаційних технологій та електротехніки та електронної інженерії. Університет транспорту міста Хошимін, вул. Во Оань, 2. Хошимін, 700000, В'єтнам