Research article Pham Ngoc Thach Medical and Pharmaceutical Journal

DOI: 10.59715/pntjmp.5.2.12

ChatCVHT: An academic advising chatbot based on semantic retrieval, with topic routing and confidence threshold calibration

Nguyen Trong Hien[1], Phung Quang Vinh[2], Le Thien Khiem[3], Dang Bao Dang[4], Nguyen Minh Tuan[5]

[1]Department of Biostatistics and Informatics, Faculty of Public Health, Pham Ngoc Thach University of Medicine, Ho Chi Minh city.

[2]Department of Environmental and Occupational Health, Faculty of Public Health, Pham Ngoc Thach University of Medicine, Ho Chi Minh city.

[3]Department of Environmental and Occupational Health, Faculty of Public Health, Pham Ngoc Thach University of Medicine, Ho Chi Minh city.

[4]Department of Organization and Health Management, Faculty of Public Health, Pham Ngoc Thach University of Medicine, Ho Chi Minh city.

[5]Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Ho Chi Minh city. (minhtuan@ptit.edu.vn)

Abstract

Academic advisors in higher education must respond to a high volume of student questions about regulations, training policies, and study planning, many of which share the same intent but differ in wording. While chatbots can scale academic support, a key risk is providing incorrect answers when queries are ambiguous or weakly supported by available evidence. We developed ChatCVHT, an academic advising chatbot that combines semantic retrieval with topic-based routing using a knowledge base of 748 question–answer pairs across eight topics. The system separates topic classification from document retrieval and introduces a confidence-based decision layer that jointly considers similarity scores, the score margin between top candidates, and predicted topic confidence to decide whether to answer or request clarification. In our experiments, multilingual-e5-small achieved stable retrieval performance (Recall@10 = 0.9782; MRR@10 = 0.8841), and multilingual-e5-small with a Logistic Regression classifier (L2 regularization, C = 3) reached Macro-F1 = 0.982 and Accuracy = 0.9853 for topic classification (5-fold cross-validation). When integrated end-to-end, the decision layer withheld responses for ~7% of queries to prioritize clarification under uncertainty, while maintaining Recall@10 = 0.916 and MRR@10 = 0.8418. Overall, ChatCVHT adopts a conservative strategy that balances coverage and reliability and supports safer deployment of academic advising chatbots where factual accuracy is critical.

Keywords: Chatbot; academic advising; semantic retrieval; sentence embedding; E5; topic classification; FAISS; confidence-threshold calibration; CLARIFY.

Received: 22/12/2025

Revised: 19/03/2026

Accepted: 20/04/2026

Author contact:
Nguyen Trong Hien

Email: hiennt@pnt.edu.vn

Phone: +84 939205330

1. INTRODUCTION

In higher education, academic advisors play a key role in helping students understand academic regulations and training policies, plan study pathways, and address emerging academic concerns. As enrollment increases, advisors receive more recurring questions, often expressed in many different ways. Handling these queries one by one is time-consuming and makes timely support harder. At the same time, institutional sources (websites, handbooks, policy documents) are not designed for interactive consultation, so students may struggle to quickly find a specific rule when they need it. Chatbots

have therefore been explored as a support tool to provide basic advising information at any time and to reduce repetitive workload. Prior reviews also note their widespread use in universities for academic support, admissions guidance, and learning assistance [1, 2]. In practice, their usefulness depends heavily on the quality of the underlying knowledge base and on the safeguards in place to ensure reliable responses.

Several design options are available, each balancing answer flexibility against information safety in different ways. One approach lets the chatbot generate answers using a large language model; without strict control of knowledge sources, this may lead to hallucinated content or extrapolation beyond formal regulations. Retrieval Augmented Generation (RAG) aims to improve factual correctness and verifiability by grounding the generation model in retrieved evidence [3, 4]. Alternatively, FAQ-style question–answering systems can achieve strong performance through semantic retrieval, where questions and answers are encoded as vectors and the system retrieves the nearest candidates under a similarity measure. Transformer-based sentence representation models have been shown to be effective across a range of retrieval tasks [5], and more recent embedding models leverage contrastive pre-training to improve multilingual representations [6]. To deploy semantic retrieval at scale with low response latency, systems commonly use Approximate Nearest Neighbor (ANN) search and appropriate indexing structures; HNSW is a commonly used ANN method with a favorable balance between retrieval quality and query speed [7]. Vector search libraries such as FAISS provide optimized implementations for various indexing scenarios and enable scaling to large datasets [8]. In Vietnam, educational chatbots have been studied and deployed through multiple lines of work, including knowledge-based approaches (e.g., ontology-based systems) and Vietnamese language processing models for tasks such as classification and intent understanding [9, 10].

However, in academic FAQ systems, the practical challenge lies not only in retrieval quality but also in safe operation. The system must avoid providing incorrect answers when questions are ambiguous, lack context, or when the retrieved evidence is insufficiently strong. This study develops an academic advising chatbot for the Faculty of Public Health, Pham Ngoc Thach University of Medicine, based on a knowledge base of 748 question-answer pairs organized into eight topics. While this dataset serves as a testbed for system development and will be expanded during real-world deployment (e.g., using structured data from official sources such as hierarchical JSON maintained by university websites and training offices), the current system is designed as a semantic retrieval pipeline integrated with topic-based routing.

To address these requirements, we propose ChatCVHT, which combines semantic retrieval with topic routing and a conservative confidence gate that triggers CLARIFY when a query is ambiguous or weakly supported by evidence. Specifically, we (1) benchmark multilingual sentence embedding models for retrieval using Recall@K and MRR@K; (2) train a lightweight topic classifier under an embedding–classifier head framework and select it using Macro-F1 and Accuracy via 5-fold cross-validation; and (3) apply confidence-threshold calibration (similarity score, margin between top candidates, and predicted topic confidence) to decide whether to answer or request user

clarification in ambiguous cases. Operational effectiveness is reported using Recall, MRR, and the CLARIFY rate as an indicator of safety-oriented abstention.

2. METHODS

2.1. Data collection and topic labeling

The dataset was collected at Pham Ngoc Thach University of Medicine and consists of 748 question–answer pairs. The questions were extracted from students' real-world inquiries. The answers were compiled from official regulations, faculty guidance documents, the student handbook, and information available on the university website. To enable faster querying, we organized the 748 question–answer pairs into eight advising topics: university overview (GIOI THIEU), faculty-related information (CAC KHOA), information from the Office of Undergraduate Academic Affairs (QLDT DH), specialty-related information (CHUYEN KHOA), Information regarding master's and doctoral students (CH-NCS), program learning outcomes (CDR NGANH), information from the Faculty of Public Health (KHOA YTCC), and other information (KHAC). The questions and answers were standardized by converting to lowercase, removing repeated punctuation, unnecessary characters, and redundant whitespace. The text was then tokenized to support subsequent representation steps and the retrieval and classification experiments. While the current dataset serves as a testbed limited to internal questions, future expansions will incorporate broader sources such as the university website, Student Affairs Handbook, and Training Management Office documents to enhance coverage of organizational structures, personnel policies, and functional departments. To improve generalization and maintainability, we recommend transitioning from discrete Q&A pairs to a hierarchical data structure (e.g., tree or multilevel JSON), starting with core entities like "Organizational Structure", "Training Fields", "Student Policies", and "Healthcare Resources". This allows defining fixed "information variables" (e.g., scholarship levels) with dynamic values, facilitating easy updates and scalability.

2.2. System architecture

ChatCVHT is designed as a retrieval-based semantic QA system. Figure 1 illustrates the architecture: (i) text normalization and the creation of two embedding representations for classification and retrieval; (ii) topic classification to estimate p(t | q) and top-N retrieval using FAISS; (iii) reranking using a combined score that integrates semantic similarity and topic confidence; and (iv) a decision module that either returns an answer or triggers the CLARIFY mechanism based on confidence thresholds (similarity, margin, and topic confidence). The system then returns the answer corresponding to the most appropriate question–answer entry. The pipeline is intended to be maintainable in a university setting where policies change; updates can be applied by refreshing the knowledge base and index, without full retraining, as detailed in Section 2.6.

We use two embedding spaces—one for topic classification and one for retrieval—because the objectives differ: classification benefits from separability in the label space, whereas retrieval benefits from fine-grained semantic matching between user queries and candidate FAQs. In practice, decoupling these representations helps reduce interference between tasks and keeps the deployed components lightweight and interpretable, which is important for an early-stage prototype under limited computational resources.

2.3. Semantic representation and FAISS-based retrieval

Each question in the knowledge base is mapped to a semantic vector using a Transformer-based sentence embedding model. In this study, the embedding models include intfloat/multilingual-e5-small (E5), pre-trained with a semantic search objective [6]; bkai-foundation-models/Vietnamese-bi-encoder (VBE), optimized for Vietnamese [14]; and MiniLM-based Sentence-Transformers models such as paraphrase-multilingual-MiniLM-L12-v2 (mMiniLM) and all-MiniLM-L6-v2 (MiniLM), which are widely used for sentence representation and semantic retrieval [5, 12].

Given an input query q, the system generates a query vector using the same model (or the selected retrieval model). The similarity between the query and the i-th question is computed using cosine similarity:

$$S(q, d_i) = \frac{q.d_i}{\|q\|\|d_i\|} \quad (1)$$

The set of vectors {di} is indexed with FAISS to accelerate nearest-neighbor search at scale [8]. At query time, the system retrieves the top-K candidates with the highest similarity under s(·) and uses these candidates for subsequent reranking and the chatbot's final decision-making.

Because the knowledge base in this study is relatively small, we use an exact FAISS index (IndexFlatIP) to keep ranking behavior stable. Embeddings are L2-normalized, and cosine similarity is implemented via inner product on normalized vectors. At the current scale, exact search is fast enough and keeps the retrieval setup straightforward to reproduce.

2.4. Topic classification

The topic classification task is formulated under an embedding–classifier head framework with $C = 8$ classes. Each text is mapped to a semantic vector using a sentence embedding model and then fed into a shallow classifier to predict the topic label $\hat{y} \in \{1,...,C\}$. This design separates the roles of representation and decision, enabling a direct assessment of the embeddings' degree of linear separability in the label space. The classifier heads are denoted as follows: LR-C1 / LR-C3 (Logistic Regression with L2 regularization and C ∈ {1,3}), SVC-cal (LinearSVC with probability calibration), KNN-cos (KNN with cosine distance), and GNB (Gaussian Naive Bayes). All configurations are evaluated using 5-fold cross-validation on the dataset with N = 748 samples and C = 8 classes. Two primary metrics are reported: (i) MacroF1 (the class-averaged F1 score, mitigating bias due to class imbalance), (ii) Accuracy (the proportion of correct predictions). Results in the table are presented as mean ± std over the five folds. Separate embedding spaces for classification and retrieval reduce noise; Logistic Regression (LR-C3) is prioritized for its simplicity, fast inference, and stable performance on small datasets, which fits deployment scenarios with limited computational resources.

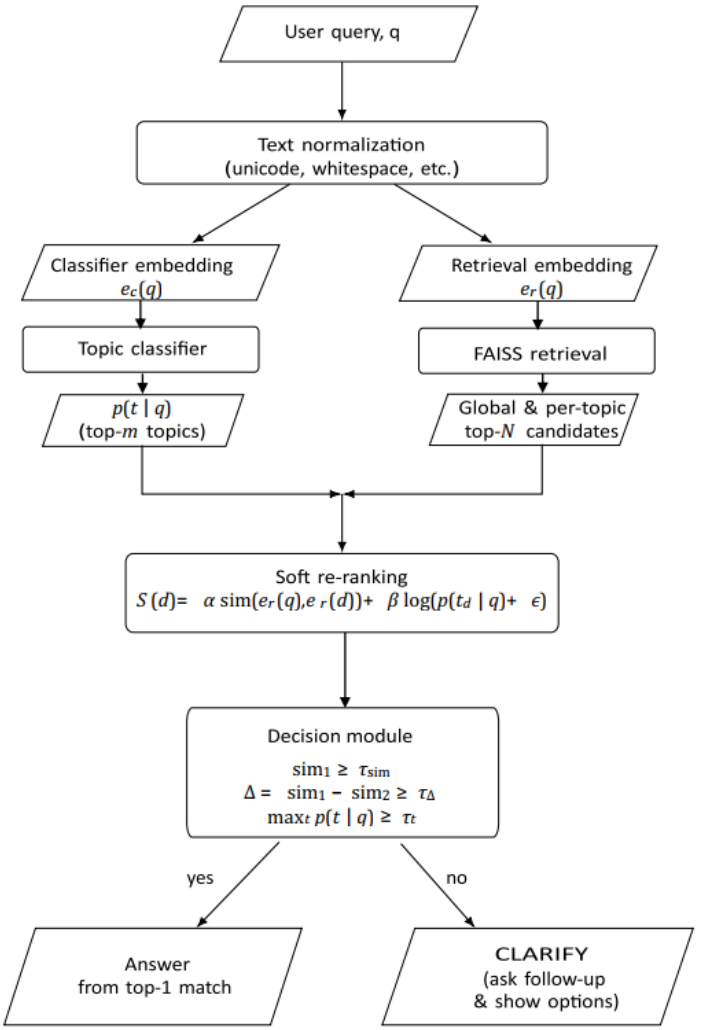

Figure 1. Vertical ChatCVHT pipeline. Rounded rectangles denote processing modules (text normalization, topic classifier, FAISS retrieval, soft reranking, and decision module), while parallelograms denote data objects and intermediate representations (e.g., user query, embeddings, topic probabilities, retrieved candidates, and outputs).

2.5. Confidence-threshold calibration and the CLARIFY mechanism

A retrieval-based chatbot may produce incorrect answers when queries are ambiguous, lack context, or fall outside the coverage of the knowledge base. To reduce this risk the system applies confidence-threshold calibration and returns an answer only when three conditions are simultaneously satisfied: (i) the top-1 candidate similarity score s_1 is at least the threshold τ_{sim} (SIM_THR); (ii) the margin $\Delta = s_1 - s_2$ between the top-1 and top2 candidates is at least τ_Δ (MARGIN_THR), ensuring that the top-1 candidate is sufficiently dominant; and (iii) the maximum predicted topic probability $\max_t p(t \mid q)$ is at least τ_t (TOPIC_THR). If any condition is not met, the system does not answer directly; instead, it triggers the *CLARIFY* mechanism to request additional user information (e.g., major, cohort, or program) and may suggest a set of relevant options to facilitate query disambiguation. This mechanism encourages conservative behavior under uncertainty and reduces the risk of providing incorrect information.

User-facing clarification. When CLARIFY is triggered, the system explicitly informs the user that additional context is needed

(rather than returning a potentially incorrect answer) and asks a short follow-up question (e.g., cohort/program/major). In the prototype interface, CLARIFY is accompanied by suggested options (quick-reply buttons) to reduce user effort and mitigate the perception of system failure during conservative abstention.

2.6. Knowledge Update Mechanism

To address frequent policy changes in university environments without requiring full retraining, the system supports flexible updates via FAISS index refresh: (1) Fetch new or updated data from official sources (e.g., university website, Student Affairs Handbook); (2) Encode and insert new vectors into the index; (3) Validate with sample queries to ensure retrieval quality.

For future scalability, the architecture allows optional integration of large language models (LLMs, e.g., GPT-3.5 or later versions) as a post-retrieval synthesizer. In such a hybrid setup, LLMs would only be invoked after the system passes all confidence thresholds (SIM_THR, MARGIN_THR, TOPIC_THR) and CLARIFY is not triggered. The LLM could then rephrase retrieved answers naturally or access dynamic variables in hierarchical JSON structures (e.g., summarizing current scholarship levels from structured data). This controlled integration would enhance response fluency while CLARIFY continues to serve as a safety gate, preventing LLM invocation on ambiguous or low-confidence queries and thereby mitigating hallucination risks. The current implementation remains purely retrieval-based to prioritize factual accuracy and deployment safety.

Notably, updating the knowledge base does not require retraining the embedding model; it only requires encoding new or updated entries and refreshing the FAISS index. If topic distributions change substantially during real deployment, the lightweight topic classifier can be re-trained periodically using newly collected labeled queries.

2.7. Evaluation protocol and threshold selection

We evaluate three components of ChatCVHT: (i) dense retrieval, (ii) topic classification, and (iii) the end-to-end pipeline with the CLARIFY decision layer. Topic classification is assessed using 5-fold cross-validation on the full dataset (N = 748, C = 8), reporting Macro-F1 and Accuracy as mean ± std across folds. The complete pipeline is evaluated on an annotated query set with N_{eval} = 643, reporting Recall@10, MRR@10, and CLARIFY_rate (the fraction of queries routed to clarification).

Threshold selection is treated as a safety–coverage trade-off rather than pure accuracy maximization. We tune (τ_{sim}, τ_{Δ}, τ_t) on a held-out development split by choosing settings that maintain strong ranking quality (MRR@10) while limiting risky low-evidence answers, and we report the chosen thresholds together with end-to-end results in Table 4. This design choice reflects the safety-critical nature of academic advising, where abstaining and asking for clarification is preferable to overconfident incorrect responses.

2.8. User Interface and Prototype Deployment

To support experimental validation of the ChatCVHT development workflow, we deploy a lightweight web-based prototype with a chat-style interface as an end-to-end testbed.

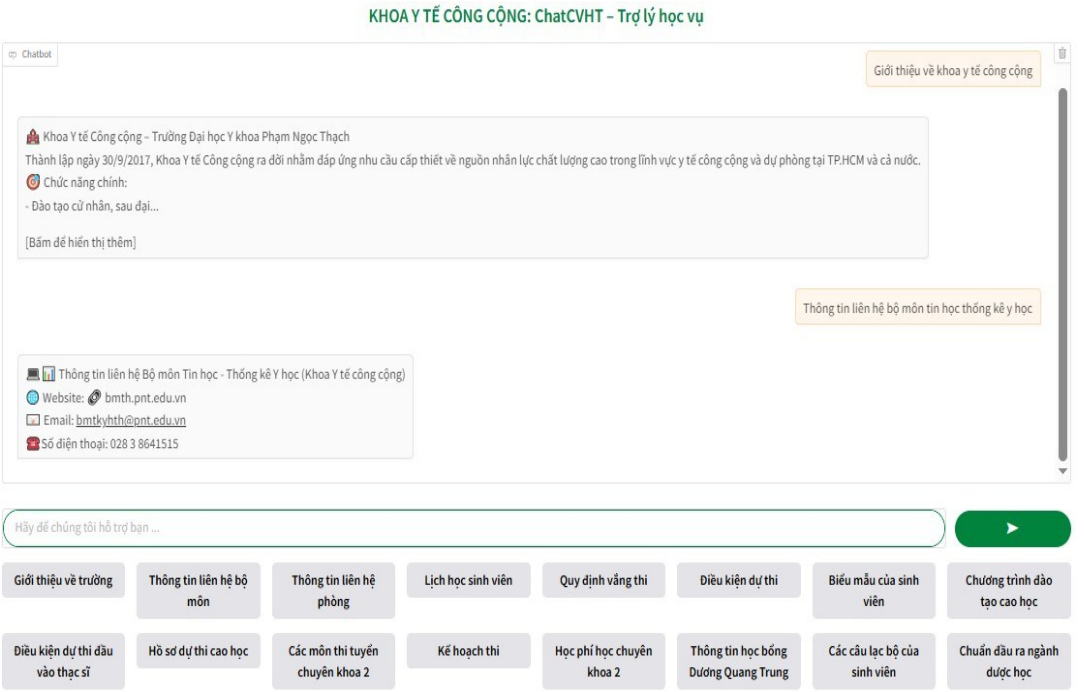

Figure 2. Web-based prototype interface of ChatCVHT used as an experimental testbed. The UI supports chat-style interaction, quick-reply topic buttons, and structured answer cards; the system triggers a CLARIFY follow-up when confidence thresholds are not met.

The goal of this prototype is to evaluate the complete chatbot pipeline under realistic interactive usage (rather than to deliver a finalized production system), including free-text question input, knowledge retrieval, response presentation, and uncertainty handling. Users submit questions through an input box, and the system returns answers as structured message cards (e.g., bullet points, contact details, and hyperlinks) to improve readability for administrative and policy-related information. To reduce typing effort and mitigate ambiguity, the interface provides topic shortcuts and frequently used options as quick-reply buttons. The prototype uses the same knowledge sources described in the dataset section (i.e., the curated Q&A set and official academic-advising information), which are indexed for retrieval. On the backend, the UI invokes the ChatCVHT pipeline (topic routing, FAISS-based retrieval, reranking, and confidence-threshold checks). If confidence constraints are not satisfied, the system triggers the CLARIFY interaction and presents follow-up prompts or suggested options to disambiguate the query before returning an answer, thereby prioritizing safe clarification over potentially incorrect responses.

3. RESULTS

3.1. Distribution of advising topics

Table 1 presents the distribution of question–answer pairs across the eight topics. The data show that questions are primarily concentrated in the faculties and units category (22.9%) and the specialties category (21.0%), which together account for 43.9%. The general introduction category (14.7%) and the Information regarding master's and doctoral students (13.9%) also represent substantial proportions (28.6%). The remaining topics have lower shares (each below 11%), yet they still contribute to ensuring that the dataset covers diverse information needs, including undergraduate academic administration, information from the Faculty of Public Health, and program learning outcomes.

Table 1. Distribution of question–answer pairs by advising topic (n = 748)

Category	Count	Percentage (%)
CAC KHOA	171	22.9
CHUYEN KHOA	157	21.0
GIOI THIEU	110	14.7
CH – NCS	104	13.9
KHAC	82	11.0
QLDT DH	49	6.6
KHOA YTCC	42	5.6
CDR NGANH	33	4.4

The dataset exhibits moderate diversity, with an average query length of 7.82 words (SD = 2.53; range: 1–19 words). Semantic overlap across topics is non-trivial (e.g., "scholarship" may appear in both Student Policies and Tuition), which can increase task difficulty and raise overfitting concerns in small datasets; however, the low 5-fold cross-validation variance (<0.01) suggests stable generalization within the testbed.

3.2. Benchmarking embedding models for retrieval

Table 2 shows that E5 consistently outperforms the compared baselines, achieving Recall@10 = 0.9782 and MRR@10 = 0.8841, which indicates that the correct entry is frequently ranked among the top candidates. In terms of preprocessing cost, VBE incurs the highest embedding time (DocEmbed(s)), whereas E5 is slower than the MiniLM-based models but provides a clear gain in retrieval accuracy.

E5's strong performance is likely attributable to its contrastive pre-training, which improves semantic matching under overlapping intents, while remaining effective in a small-data setting. For benchmarking, we use an 80/20 hold-out protocol (598/150 Q–A pairs): queries are taken from the held-out 20% set, and retrieval is performed against an index built from the remaining 80% subset.

Table 2. Benchmark results of embedding models for retrieval

Model	DocEmbed(s)	Recall@1	Recall@10	MRR@10
E5	34.598	0.8398	0.9782	0.8841
VBE	50.811	0.8274	0.9642	0.8780
mMiniLM	28.342	0.7185	0.9269	0.7874
MiniLM	17.400	0.7014	0.9067	0.7690

Note: *DocEmbed(s) denotes the time required to generate embeddings for all 748 documents in a single run.*

3.3. Topic classification results and the optimal configuration

Among all evaluated combinations, E5 + LR-C3 achieves the highest MacroF1 (0.982) and Accuracy (0.9853) under 5-fold cross-validation. The foldwise variance is small for VBE with Logistic Regression, suggesting stable generalization, while mMiniLM exhibits larger fluctuations across splits, consistent with weaker class separation in its embedding space. Logistic Regression provides the strongest and most consistent results across embeddings; in this dataset, SVC-cal offers no measurable benefit, whereas GNB and KNN-cos underperform, likely due to mismatch between their assumptions and the geometry of the embedding space. Based on the highest Macro-F1 and Accuracy, together with cross-validation stability, the system selects embedding E5 and the LR-C3 head as the optimal configuration for the topic classification module in the current version. This choice also simplifies deployment (a linear head with fast inference) and improves output consistency when integrated with subsequent components (retrieval and reranking).

High Macro-F1 (0.982) on a small dataset may raise overfitting concerns; however, the low 5-fold cross-validation variance (<0.01) suggests stable generalization within the testbed. In 5-fold CV, each fold uses approximately 80% of the data for training and 20% for testing, and the diversity in query lengths may contribute to class separability.

Table 3. Topic classification results under the "Embedding → Classifier head" framework (5-fold CV, N = 748, C = 8)

Emb.	Head	Macro-F1	Accuracy
E5	LR-C3	0.982 ± 0.017	0.985 ± 0.012
E5	LR-C1	0.979 ± 0.014	0.983 ± 0.010
E5	SVC-cal	0.978 ± 0.013	0.980 ± 0.011
VBE	LR-C1	0.956 ± 0.005	0.965 ± 0.003
VBE	LR-C3	0.953 ± 0.005	0.963 ± 0.006
VBE	SVC-cal	0.951 ± 0.011	0.963 ± 0.008
mMiniLM	LR-C1	0.942 ± 0.036	0.949 ± 0.030
mMiniLM	LR-C3	0.939 ± 0.031	0.944 ± 0.028

Note: *Emb.=embedding; LR-C1/LR-C3: Logistic Regression (L2) with parameter C; SVC-cal: LinearSVC with probability calibration; VBE: Vietnamese bi-encoder; mMiniLM: multilingual SBERT with a MiniLM backbone.*

3.4. Results of the complete chatbot pipeline and chatbot configuration

Table 4 summarizes the evaluation results of the chatbot pipeline under representative configurations of topic classification and retrieval models. The reported metrics include Recall@10, MRR@10, and the *CLARIFY* activation rate. The table also lists the corresponding calibration thresholds (SIM_THR, MARGIN_THR, and TOPIC_THR). The evaluation set contains $N_{eval} = 643$ samples.

Our design analysis suggests that topic routing and a margin-based rule help stabilize ranking and promote conservative CLARIFY behavior under uncertainty. Prior RAG best-practice studies highlight that additional modules (e.g., HyDE-style generation) can

increase latency and that factual errors may still occur under imperfect grounding, motivating clarification-first strategies [16].

The E5 + LR_l2_C3 configuration with E5 retrieval achieves Recall@10 = 0.9160, MRR@10 = 0.8418, and a CLARIFY_rate of 0.0731 (with N_{eval} = 643). When the retrieval model is replaced with VBE, performance remains nearly comparable (MRR@10 = 0.8438, CLARIFY_rate = 0.0653). Because VBE retrieval yields comparable MRR@10 to E5 in our setting, using E5 for both topic classification and retrieval simplifies the system without sacrificing overall performance.

Table 4. Chatbot pipeline results under representative configurations (Neval = 643)

ClsEmb	Head	RetEmb	SIM_THR	MARG_THR	TOPIC_THR	R@10	MRR@10	CLAR
E5	LR-C3	E5	0.8838	0.0003	0.35	0.916	0.8418	0.0731
E5	LR-C3	VBE	0.6022	0.0038	0.35	0.916	0.8438	0.0653
E5	LR-C3	mMiniLM	0.7373	0.0009	0.35	0.8896	0.7690	0.0653
E5	LR-C3	MiniLM	0.7248	0.0000	0.35	0.874	0.7591	0.0902

Note. *ClsEmb: embedding model used for topic classification; RetEmb: embedding model used for retrieval; Head: classifier head. SIM_THR/MARG_THR/TOPIC_THR: calibration thresholds used for routing decisions and CLARIFY activation (corresponding to the similarity threshold, the margin threshold, and the topic-confidence threshold). R@10: Recall@10; MRR@10: Mean Reciprocal Rank@10; CLAR: proportion of queries that trigger CLARIFY ; LR-C3=LR_l2_C3.*

3.5. Error analysis and qualitative examples

To better understand the residual errors beyond aggregate Recall/MRR, we analyze system decisions on the evaluation set (N_{eval} = 643) by separating (i) answered queries and (ii) queries routed to *CLARIFY*. Table 5 summarizes the outcomes. The system answered 601 queries, among which 514 were correct, yielding an accuracy of 0.8552 on answered cases. The remaining 42 queries (CLARIFY_rate = 0.0653) were withheld to reduce the risk of incorrect responses.

Table 5. Error analysis summary on the evaluation set (Neval = 643)

Decision	Outcome type	Count	Percent (%)
ANSWER	Correct	514	79.94
ANSWER	Wrong answer: ranking error	75	11.66
ANSWER	Wrong answer: retrieval miss	12	1.87
CLARIFY	Over-conservative CLARIFY (gold@1)	28	4.35
CLARIFY	CLARIFY despite gold in top-*K*	10	1.56
CLARIFY	CLARIFY (low evidence / ambiguous)	4	0.62

A content-level evaluation (beyond topic matching) is planned via expert review on a sampled set of user queries, including out-of-scope or complex cases (e.g., “international scholarships 2026”), to assess factual correctness (e.g., scholarship levels) and characterize failure modes.

Answered errors. Among answered mistakes, the dominant failure mode is ranking error (75 cases, 11.66%), where the gold item is still present in the retrieved top-K list but is not selected as top-1. This pattern indicates that retrieval quality is generally sufficient, while the final selection can be sensitive to near-duplicate or semantically adjacent FAQs (e.g., pairs such as "mission" vs. "vision", or closely related identity questions). In contrast, retrieval miss is relatively rare (12 cases, 1.87%) and mainly appears in off-domain or conversational queries (e.g., greetings or social chit-chat), which are outside the coverage of the current academic FAQ knowledge base.

CLARIFY behavior and causes. The majority of CLARIFY cases are *over-conservative* (28 cases, 4.35%), where the gold answer is already ranked at top-1 but the guardrails still reject answering. This suggests that the confidence thresholds trade a small amount of coverage for improved safety. Importantly, only 4 queries (0.62%) are routed to CLARIFY due to genuinely weak evidence or ambiguity, indicating that the mechanism is selective rather than overly restrictive. We further inspect which confidence constraints most frequently trigger CLARIFY. As shown in Table 6, the margin constraint is the primary driver, while topic confidence does not act as an active bottleneck in this configuration.

Table 6. Counts of confidence constraints associated with CLARIFY decisions

Constraint	Count	Share among CLARIFY (%)
Similarity below threshold	12	28.57
Margin below threshold	31	73.81
Topic confidence below threshold	0	0.00
Any CLARIFY	42	100.00

Qualitative examples. Representative failure cases are consistent with the quantitative analysis: (i) *retrieval-miss* tends to occur for out-of-scope conversational inputs (e.g., "Ai tạo ra bạn"), motivating a lightweight intent filter or an explicit "out-of-domain" response policy; (ii) *ranking errors* often involve semantically close FAQ pairs (e.g., "Bạn là ai" vs. "Ai tạo ra bạn", or "Sứ mạng" vs. "Tầm nhìn"), suggesting that deduplication/canonicalization of near-duplicate questions or a stronger reranking strategy could further reduce this class of errors; and (iii) *over-conservative CLARIFY* occurs when the top-1 candidate is correct but violates SIM/MARGIN thresholds, indicating that adaptive thresholds or class-conditional calibration may increase coverage without materially increasing risk.

4. DISCUSSION

In the end-to-end evaluation, ChatCVHT attains Recall@10 = 0.916 and MRR@10 = 0.8418, with clarification triggered for only 6–7% of queries. Compared with prior work reporting high recall under a fixed threshold, our design emphasizes operational safety by explicitly routing low-confidence cases to *CLARIFY* rather than forcing a potentially incorrect response. In contrast, Holis *et al.* [11] report Recall = 96.3% with a threshold = 0.5, but an Accuracy = 79.2% (with no mechanism specified). A key practical distinction of the proposed system is the addition of a safety decision layer (triggering *CLARIFY*) to reduce the risk of incorrect responses. When queries are ambiguous or confidence scores are insufficiently high, the system accepts a

partial trade-off in coverage to prioritize response safety and reliability.

The experiments also show that E5 yields higher retrieval quality than general-purpose Sentence-BERT variants (e.g., multilingual MiniLM). We also note that E5 supports prefix-based preprocessing using query: and passage: enabling the embedding space to learn the intended roles of query and candidate for retrieval [15]. Conversely, SBERT was proposed as a general sentence embedding framework based on a Siamese/bi-encoder architecture; it is broadly applicable but does not necessarily optimize top-k ranking for every domain [5].

The topic classification results, achieving Macro-F1 = 0.982 with the E5 + Logistic Regression configuration, indicate that high-quality embeddings yield a feature space with strong class separability. This, in turn, enables a linear classifier to approach near-optimal performance. The results also reinforce the role of *topic routing*: (i) reducing the risk of off-domain retrieval as the knowledge base grows, and (ii) mitigating noise from cross-topic FAQs with overlapping vocabulary.

The SIM/MARGIN/TOPIC thresholds provide a conservative decision rule: when evidence is weak, the system requests clarification instead of answering. The observed CLARIFY_rate (6–7%) suggests that this safeguard is selective rather than overly restrictive, while still filtering ambiguous queries.

Recent surveys indicate that educational chatbots are widely used for academic advising, learner support, and reducing repetitive workload, while also emphasizing challenges related to reliability and the risk of misinformation [1, 2]. Against this backdrop, the proposed system's *CLARIFY* mechanism and confidence thresholds constitute a pragmatic design choice to improve trustworthiness in academic advising chatbot deployment.

Given its query-time efficiency and system-level optimizations [8], FAISS is a suitable choice for ANN-based vector search at scale. If the system is extended toward RAG, as suggested by recent surveys [2, 3, 4], the calibration layer and *CLARIFY* mechanism should still be retained as a "safety valve" prior to triggering generation [15].

This work focuses on an early-stage, retrieval-based prototype grounded in curated institutional FAQs, where conservative safety behavior is a primary design goal. Due to practical constraints in computational resources and the need to keep the revision minimal at this stage, we do not include additional end-to-end baselines such as LLM-only or BM25-RAG comparisons in the current manuscript. We instead position these comparisons as planned follow-up experiments when larger-scale deployment data and stronger experimental resources become available.

A pure retrieval system with CLARIFY remains a reasonable choice in safety-critical contexts like medical education, where avoiding misinformation (e.g., incorrect scholarship details) outweighs RAG's flexibility in synthesizing responses. This trade-off prioritizes verifiability over expressiveness, as RAG's generative component risks hallucinations despite retrieval grounding. However, our approach currently underutilizes LLMs' synthesis capabilities, limiting natural rephrasing or context-adaptive answers. Future integration could position ChatCVHT as a front/back-end layer for RAG: e.g., CLARIFY as a pre-generation gate (only activate LLM if confidence high) or post-verification (check generated output against retrieved sources), fostering a hybrid model that balances safety with

RAG's advantages for long-term objectivity and alignment with current research trends.

Large language models (e.g., GPT-3.5) can improve response drafting by first inferring user intent, then pulling relevant structured signals from hierarchical variables (such as a "scholarship level" field in JSON), and finally producing fluent, natural-sounding replies. This adds practical advising value beyond retrieval alone, while the CLARIFY mechanism is still retained to preserve reliability.

We select E5 + LR-C3 as the best overall configuration because it delivers strong retrieval/classification results and remains lightweight to deploy. In the end-to-end evaluation, the CLARIFY decision layer withheld answers for about 7% of queries, reflecting a cautious, safety-oriented stance when the system is uncertain.

Limitations and future work

This study is limited by the size and institutional scope of the current knowledge base (748 question–answer pairs from a single faculty), which constrains external generalizability. The evaluation in this manuscript emphasizes automatic retrieval and classification metrics and an end-to-end safety indicator (CLARIFY rate); it does not yet include large-scale user-centered evaluation of answer usefulness and factual correctness under diverse real-world interactions. In addition, given that this work represents an early-stage prototype developed under limited experimental resources, we do not report broader model comparisons (e.g., LLM-only, BM25-RAG, or stronger reranking variants) in the current revision. These comparisons are part of our planned next-stage research when deployment conditions and experimental resources allow more extensive benchmarking. In subsequent work, we will (i) expand the knowledge base using structured official sources (e.g., hierarchical JSON variables for frequently changing policies), (ii) conduct user-centered evaluation with target stakeholders (students and academic advisors/counselors) using expert review and structured rubrics to obtain objective feedback on correctness, clarity, and safety, and (iii) evaluate hybrid retrieval-and-generation settings where CLARIFY and confidence thresholds remain as safety gates to reduce the risk of hallucinated outputs.

5. CONCLUSION

This study presents ChatCVHT, an early-stage academic advising chatbot that combines semantic retrieval with topic-based routing and a confidence-based CLARIFY decision layer. Using a knowledge base of 748 question–answer pairs across eight topics, we benchmark embedding models for retrieval and evaluate lightweight topic classification configurations, selecting multilingual-e5-small for stable retrieval (Recall@10 = 0.9782; MRR@10 = 0.8841) and multilingual-e5-small with Logistic Regression (L2, C = 3) for topic classification (Macro-F1 = 0.982; Accuracy = 0.9853). In the end-to-end pipeline, confidence thresholds encourage conservative behavior on ambiguous queries by triggering clarification (~7% of cases), thereby reducing the likelihood of incorrect responses in safety-sensitive academic advising. Future work will extend evaluation with stakeholder feedback and broader model comparisons under improved experimental resources.

REFERENCES

1. L. Labadze, M. Grigolia, and A. Machaidze, "Role of AI Chatbots in Education: Systematic Literature Review," International Journal of Educational Technology in Higher Education, 2023, doi:10.1186/ s41239023-00426-1.

2. T. Debets et al., “Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts,” Computers & Education, vol. 234, Sep. 2025, 105323, doi:10.1016/j.compedu.2025.105323.
3. M. Klesel and H. F. Wittmann, “Retrieval-Augmented Generation (RAG),” Business & Information Systems Engineering, vol. 67, no. 4, pp. 551–561, 2025, doi:10.1007/s12599-025-00945-3.
4. Z. Li et al., “Retrieval-augmented generation for educational application: A systematic survey,” Computers and Education: Artificial Intelligence, vol. 8, Jun. 2025, 100417, doi:10.1016/j.caeai.2025.100417
5. N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” arXiv, 2019, doi:10.48550/arXiv.1908.10084.
6. L. Wang et al., “Text Embeddings by Weakly-Supervised Contrastive Pre-training,” arXiv, 2022, doi:10.48550/arXiv.2212.03533.
7. Y. A. Malkov and D. A. Yashunin, “Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 824–836, 2020, doi:10.1109/TPAMI.2018.2889473.
8. J. Johnson, M. Douze, and H. Jégou, “Billion-scale Similarity Search with GPUs,” IEEE Transactions on Big Data, 2019, doi:10.1109/TBDATA.2019.2921572.
9. T. T. S. Nguyen et al., “An Ontology-Based Question Answering System for University Admissions Advising,” Intelligent Automation & Soft Computing, vol. 36, no. 1, pp. 601–616, 2023, doi:10.32604/iasc.2023.032080.
10. D. Q. Nguyen and A. T. Nguyen, “PhoBERT: Pre-trained language models for Vietnamese,” arXiv, 2020, doi:10.48550/arXiv.2003.00744.
11. R. M. Holis, P. E. P. Utomo, and B. F. Hutabarat, “Semantic FAQ Chatbot Using SBERT (Sentence-BERT) and Cosine Similarity for Academic Services,” Brilliance: Research of Artificial Intelligence, vol. 5, no. 2, pp. 915–922, 2025, doi:10.47709/brilliance.v5i2.7027.
12. W. Wang et al., “MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers,” arXiv, 2020, doi:10.48550/arXiv.2002.10957.
13. D. Wang et al., “A Comprehensive Survey and Experimental Comparison of Graph-based Approximate Nearest Neighbor Search,” Proceedings of the VLDB Endowment, 14(11), 2021, doi:10.14778/3476249.3476255.
14. N. Q. Duc et al., “Towards Comprehensive Vietnamese Retrieval-Augmented Generation and Large Language Models,” arXiv, 2024, doi:10.48550/arXiv.2403.01616.
15. L. Wang et al., “Multilingual E5 Text Embeddings: A Technical Report,” arXiv, 2024, doi:10.48550/arXiv.2402.05672.
16. X. Wang et al., “Searching for Best Practices in Retrieval-Augmented Generation,” arXiv, 2024, doi:10.48550/arXiv.2407.01219.