Jurnal Testing dan Implementasi Sistem Informasi
Vol 4(1) 2026 : 1-19

IMPROVING FAQ RETRIEVAL FOR ACADEMIC REGULATIONS USING SEMANTIC EMBEDDINGS AND LLM QUESTION AUGMENTATION

Fajri Profesio Putra[1]*, I Gusti Agung Putu Mahendra[2], Agus Tedyyana[3], Muhammad Noor[4]
Department of Informatic Engineering, Politeknik Negeri Bengkalis, Bengkalis, Indonesia[123]
School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia[4]
fajri@polbeng.ac.id*

Received: 12 March 2026, Revised: 19 May 2026, Accepted: 23 May 2026
**Corresponding Author*

ABSTRACT
Academic regulations in higher education are often documented in lengthy and formal handbooks, making it difficult for students to find relevant information using everyday language. This study developed a semantic FAQ retrieval system for academic regulations using IndoSBERT and question augmentation. The FAQ corpus was constructed from official academic and internship documents, resulting in 92 FAQ entries across 33 topical categories. Seed questions were generated from category–keyword pairs and expanded using simple rule-based augmentation and FLAN-T5-based paraphrasing. The dataset was evaluated using an 80:10:10 train–validation–test split. IndoSBERT was fine-tuned with Multiple Negatives Ranking Loss under three configurations: baseline, baseline with simple augmentation, and baseline with simple plus LLM-based augmentation. Retrieval performance was measured using Recall@1, Recall@3, Recall@5, and Mean Reciprocal Rank. The best result was achieved by the simple plus LLM augmentation configuration, with Recall@1 of 0.7848, Recall@5 of 0.8987, and MRR of 0.8396. These findings show that LLM-based question augmentation improves semantic retrieval robustness while keeping answers grounded in curated academic regulations.

Keywords : *academic chatbot, FAQ retrieval, IndoSBERT, question augmentation, semantic search.*

1. Introduction

Artificial intelligence (AI) has increasingly been adopted in higher education to support academic information services, particularly for handling repetitive questions related to regulations, procedures, and student services. Universities and polytechnics manage complex administrative processes, including admission, registration, academic leave, internship, course enrollment, disciplinary rules, and graduation. These regulations are usually documented in official academic handbooks or procedural guidelines. Although such documents provide authoritative information, they are often lengthy, formal, and difficult to search using everyday student language. As a result, students frequently rely on administrative staff, academic advisors, or messaging groups to ask repeated questions, especially during critical academic periods such as registration, internship submission, and graduation preparation. (Crompton & Burke, 2023).

A key challenge in academic information access is the mismatch between the language used in official documents and the language used by students. Academic regulations are typically written using formal institutional terminology, while students often ask questions using informal expressions, abbreviations, synonyms, incomplete sentences, or typographical errors. For example, students may use terms such as such as “academic leave,” “graduation requirements,” or “industrial internship,” whereas official documents may use more formal terms such as “academic leave procedure,” “graduation eligibility requirements,” or “field practice/internship program.”. This gap reduces the effectiveness of simple keyword search and rule-based chatbot systems, which generally depend on exact lexical matching or predefined intent patterns.(Ghasemi & Shakery, 2024). Lexical retrieval methods such as TF-IDF and BM25 are widely used because they are efficient, interpretable, and effective when user queries contain terms that overlap with indexed documents. However, these methods are limited when the same intent is expressed through different wording. Dense semantic retrieval addresses this limitation by representing queries and candidate answers as vectors and ranking them based on semantic similarity rather than exact word overlap. Hybrid retrieval, which combines lexical and dense

scores, can further improve retrieval by balancing keyword precision and semantic flexibility. Nevertheless, hybrid retrieval still depends on the quality of the dense encoder and does not automatically solve domain mismatch when the encoder has not been adapted to local institutional terminology.(Xiang et al., 2025). For Indonesian academic FAQ retrieval, IndoSBERT is a suitable semantic encoder because it follows the Sentence-BERT bi-encoder paradigm while being adapted to Indonesian language representation. Compared with general multilingual sentence encoders, IndoSBERT is expected to better capture Indonesian sentence structure and local academic terminology. Compared with generative LLM-based chatbots, an IndoSBERT retrieval framework is also more controllable because it selects answers from a curated FAQ corpus rather than generating free-form responses (Lauriola et al., 2025). This is important in academic regulation services, where answer consistency and traceability to official policy are required. However, IndoSBERT still requires domain-specific fine-tuning because academic regulations contain institution-specific terms, abbreviations, and procedural expressions that may not be sufficiently represented in general pretraining data (Shaik et al., 2022).

Another important limitation is the scarcity of authentic student questions. Official academic documents mostly provide answers in the form of rules, definitions, procedures, and requirements, but they rarely include diverse real-world questions. Training a retrieval model only on formal template-based questions may lead to weak generalization when the system receives conversational, shortened, noisy, or mixed-language queries. Question augmentation can reduce this gap by expanding the query space associated with each canonical answer. Simple rule-based augmentation can introduce surface-level variations such as informal wording, shortened queries, and light typographical errors, while LLM-based paraphrasing can generate richer semantic variations with different sentence structures and registers (Chu et al., 2023).

Despite recent progress in semantic retrieval and educational chatbots, several research gaps remain. First, many academic FAQ systems still rely on rule-based or lexical matching and do not evaluate dense retrieval models against strong lexical baselines such as BM25 and TF-IDF. Second, studies that use semantic encoders often do not compare pretrained dense retrieval, hybrid lexical–dense retrieval, and fine-tuned retrieval under the same experimental setting. Third, the contribution of question augmentation, especially the combination of simple rule-based augmentation and LLM-based paraphrasing, remains underexplored in Indonesian academic-policy FAQ retrieval. Fourth, the use of IndoSBERT for regulation-grounded academic FAQ retrieval has not been sufficiently examined with ranking-based metrics such as Recall@k and Mean Reciprocal Rank (MRR) (Zhao et al., 2024).

Therefore, this study develops and evaluates an IndoSBERT-based semantic FAQ retrieval system for academic regulations. The study aims to: (1) construct a structured FAQ corpus from official academic and internship guidelines; (2) generate seed questions from category–keyword pairs; (3) expand the question set using simple rule-based augmentation and LLM-based paraphrasing; (4) fine-tune IndoSBERT using question–answer pairs; and (5) evaluate the model against BM25, TF-IDF, IndoSBERT without fine-tuning, and hybrid BM25 + IndoSBERT baselines. The main contribution of this study is to demonstrate that query-side semantic enrichment through multi-stage question augmentation can improve retrieval robustness in a low-resource, institution-specific academic regulation domain while keeping final answers grounded in curated official FAQ entries.

2. Literature Review

2.1. Lexical and Semantic FAQ Retrieval

FAQ retrieval is a specific form of information retrieval that aims to match a user query with the most relevant answer from a predefined knowledge base. In academic service systems, this task is important because students often ask repeated questions about regulations, procedures, study requirements, internships, graduation, and administrative services. Unlike open-ended question answering, FAQ retrieval relies on curated answer candidates, making it more suitable for institutional settings that require consistency with official documents.

Traditional FAQ retrieval commonly uses lexical matching methods such as TF-IDF and BM25. These methods are computationally efficient and interpretable because they rank candidate answers based on term overlap and term weighting. Lexical retrieval can perform well

when the user query contains words that closely match the indexed answer. However, its main limitation is sensitivity to vocabulary mismatch. For example, a student may ask “cuti kuliah” while the official document uses “cuti akademik,” or ask “syarat ikut yudisium” while the regulation uses more formal procedural wording. In such cases, lexical methods may fail even though the query and the answer are semantically related (Lauriola et al., 2025).

Semantic retrieval addresses this limitation by representing queries and answers as dense vectors in a shared embedding space. Instead of relying only on exact word overlap, semantic retrieval estimates meaning similarity using vector distance or cosine similarity. Sentence-BERT and its variants are widely used for this purpose because they are designed to produce sentence-level embeddings that can be compared efficiently. In the context of academic regulations, semantic retrieval is particularly relevant because students often use informal, shortened, or paraphrased questions that differ from the formal wording of official documents (Ajallouda et al., 2025).

Nevertheless, semantic retrieval is not without limitations. A pretrained encoder may not fully understand domain-specific institutional terminology, such as SKS, KRS, PRALA, yudisium, KTM, or study-program-specific procedures. Therefore, domain adaptation through fine-tuning is often required to align the embedding space with the target FAQ domain.. (Abdalgader et al., 2024).

2.2. Dense Retrieval and Domain Adaptation

Dense retrieval uses neural encoders to map textual inputs into fixed-dimensional embeddings. In a bi-encoder architecture, the query and candidate answers are encoded separately using the same encoder, and their similarity is measured using cosine similarity. This architecture is efficient because answer embeddings can be precomputed and stored in a semantic index, while only the user query needs to be encoded during inference (Silva & Barbosa, 2024).

For Indonesian academic FAQ retrieval, IndoSBERT is a suitable encoder because it is based on the Sentence-BERT paradigm and is adapted for Indonesian language representation. Compared with general multilingual sentence encoders, IndoSBERT is expected to better capture Indonesian sentence structure and institutional terminology. However, even IndoSBERT may not be optimal without domain-specific fine-tuning because academic regulations contain formal language, procedural expressions, and local institutional terms that are rarely dominant in general pretraining corpora.

Fine-tuning the encoder using question–answer pairs can improve retrieval performance by moving semantically matched questions and answers closer in the embedding space. In this study, fine-tuning is not only used to adapt the model to Indonesian academic regulation language, but also to teach the model that different question forms may refer to the same canonical answer. This is important because student questions often vary in wording, length, register, and completeness. (Si et al., 2025).

2.3. Hybrid Retrieval for Academic FAQ System

Lexical retrieval and dense retrieval have complementary strengths. Lexical methods such as BM25 and TF-IDF are effective when important keywords appear explicitly in both query and answer (Raza et al., 2025). They are also useful for institutional terms, abbreviations, or technical codes that should not be semantically “smoothed out.” In contrast, dense retrieval is better at handling paraphrases and semantic similarity, but it can sometimes retrieve conceptually related yet intentually incorrect answers (Peng et al., 2023).

Hybrid retrieval combines lexical and dense similarity scores to balance these strengths. A hybrid method can preserve the precision of keyword matching while also capturing semantic similarity beyond exact token overlap. This is particularly useful in academic regulations because many answers share the same entity but differ in intent. For example, “yudisium” may appear in answers about definition, requirements, procedure, and eligibility presented in Table.1 (Raiaan et al., 2024).

Table 1. Critical Comparison of Retrieval Approaches

Approach	Strength	Limitation	Relevance to This Study
TF-IDF Cosine	Simple, fast, interpretable, effective for exact word overlap	Weak against paraphrases, synonyms, and informal wording	Used as a lexical baseline
BM25	Strong lexical ranking and robust term weighting	Still depends on explicit token overlap	Used as the main lexical baseline
Dense Retrieval / IndoSBERT NoFT	Captures semantic similarity beyond keywords	May underperform without domain adaptation	Used to test pretrained semantic retrieval
Hybrid BM25 + IndoSBERT	Combines lexical precision and semantic flexibility	Requires score normalization and λ tuning	Used to evaluate lexical–dense complementarity
Fine-Tuned IndoSBERT	Adapts embeddings to academic FAQ domain	Requires labeled Q–A data and careful evaluation	Used as the proposed semantic retriever
Fine-Tuned IndoSBERT + Augmentation	Improves robustness to varied question forms	Risk of semantic drift if augmentation is noisy	Used as the proposed enhanced model

This comparison shows that no single retrieval method is sufficient for all academic FAQ scenarios. Lexical baselines are important for measuring keyword-based retrieval strength, dense retrieval is important for semantic generalization, and hybrid retrieval provides a stronger baseline for evaluating whether fine-tuned IndoSBERT contributes additional value.

2.4. Question Augmentation and Its Risks

Institutional FAQ datasets are usually small and linguistically homogeneous. Most answers are derived from formal documents, while real student questions are often informal, incomplete, abbreviated, or noisy. This mismatch creates a distribution gap between training data and real-world user input. Question augmentation is one strategy to reduce this gap by generating multiple question variants for the same canonical answer.

Simple rule-based augmentation can produce surface-level variations such as informal phrasing, shortened forms, and light typographical errors. This approach is easy to control and has low risk of changing the original meaning. However, its diversity is limited because it mainly modifies the surface form rather than introducing deeper semantic variation (Cohen et al., 2024).

LLM-based augmentation can generate richer paraphrases with different sentence structures, registers, and word choices. This can help the model learn that multiple forms of a question may point to the same answer. However, LLM-based augmentation also introduces risks. The generated question may drift away from the original intent, become too generic, include unsupported assumptions, or produce unnatural language. Therefore, LLM-generated paraphrases should not be accepted automatically. They require filtering, deduplication, and semantic relevance checking before being added to the training corpus presented in Table.2 (Sheikholeslami et al., 2025).

Table 2. Question Augmentation Strategies and Risks

Augmentation Strategy	Example	Benefit	Risk	Mitigation
Informal phrasing	"What are the graduation requirements?" → "What requirements do I need for graduation?"	Improves robustness to conversational queries	Limited semantic diversity	Combine with other augmentation types
Short query generation	"What are the graduation requirements ?"	Simulates brief chatbot-style input	Query may become underspecified	Keep answer label only if

	→ "Graduation requirements?"			intent remains clear
Typo/noisy query	"requirement" → "requirment"	Improves tolerance to spelling errors	Excessive noise may reduce clarity	Use only light typo augmentation
Keyword-style query	"What is the procedure for academic leave?" → "academic leave procedure"	Simulates search-like input	May lose question intent	Preserve core intent keyword
LLM paraphrasing	"If I want to join graduation clearance, what requirements must I meet?"	Adds semantic and syntactic diversity	Semantic drift or hallucinated assumptions	Manual/semi-automatic filtering and deduplication

2.5 FLAN-T5-Based Question Paraphrasing

FLAN-T5-Base was selected for LLM-based question augmentation because it follows a text-to-text encoder–decoder architecture and has been instruction-tuned to follow natural language prompts. This makes it suitable for controlled paraphrasing tasks, such as rewriting a formal question into several semantically equivalent variants.

The selection of FLAN-T5-Base is motivated by four considerations. First, its text-to-text formulation is appropriate for question paraphrasing because both the input and output are textual sequences. Second, instruction tuning improves its ability to follow prompts such as "rewrite this question in a conversational style" or "generate three paraphrases with the same meaning." Third, the base-scale model is relatively lighter than larger LLMs, making it more feasible for experiments in Google Colab or limited computational environments. Fourth, it can generate diverse paraphrases without requiring an external paid API, which improves reproducibility.

However, FLAN-T5-based augmentation still requires careful filtering. The model may produce repetitive, generic, or semantically shifted outputs, especially when the prompt is not sufficiently constrained or when the target language/domain is underrepresented. Therefore, in this study, FLAN-T5 is used only as a paraphrase generator, not as an answer generator. The final chatbot remains retrieval-based, meaning that all responses are selected from curated FAQ answers rather than generated freely by the LLM presented in Table.3.

Table 3. Rationale for Selecting FLAN-T5-Base

Criterion	Justification
Architecture	Encoder–decoder text-to-text model suitable for paraphrasing
Instruction tuning	Better ability to follow natural-language paraphrasing prompts
Computational feasibility	Base-size model is more realistic for Colab-scale experiments
Reproducibility	Can be run locally without relying on closed external APIs
Role in this study	Used only for question augmentation, not for generating final answers
Risk control	Outputs are filtered to remove duplicates, irrelevant questions, and semantic drift

2.6. Evaluation Metrics in Dense Retrieval

Core intention behind evaluating semantic retrieval arrangement revolve around measuring extent to which relevant response succeed appearing near apex position inside ranked candidate enumeration produced by system. Unlike classification paradigm, which fixate on single-label prediction accuracy, retrieval framework engender ordered candidate list predicated on semantic similarity scoring mechanism(Fang et al., 2024). Consequently, evaluation process inherently depend on ranking-oriented metric. Recall@k measure proportion of query for which correct answer manifest within top-k retrieved result. In formal term, Recall@k stand as defined as:

$$Recall@k = \frac{1}{N}\sum_{i=1}^{N} 1(Rank_i \leq k) \quad (1)$$

Recall@1 efficaciously represent top-ranked retrieval accuracy, reflecting frequency with which system assign correct answer to paramount position. Meanwhile, Recall@3 and Recall@5

assess whether correct response serve as included among top three or top five candidate, respectively aspect that constitutes singularly germane for chatbot system and interactive assistance application (Salemi & Zamani, 2024). In addition to Recall-based metric, study also deploy Mean Reciprocal Rank (MRR), which incorporate exact rank position of foremost correct answer into calculation(Kartiyanta et al., 2025). MRR is defined as:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}^{i}} \quad (2)$$

Index assign greater value when correct response emerge at earlier rank position. As illustration, correct item occupying rank one yield 1.0, rank two yield 0.5, continuing in reciprocal manner. Consequently, MRR do not merely inspect presence of correct retrieval, but also rapidity with which item that surface within ordered candidate sequence (Datta et al., 2026).

3. Research Methods

This study developed a semantic FAQ retrieval system for POLBENG academic regulations using IndoSBERT and multi-stage question augmentation. The task was formulated as retrieval-based question answering, where the system selected the most relevant answer from a curated FAQ corpus rather than generating free-form responses. This design was selected to maintain consistency with official academic regulations and reduce the risk of unsupported or hallucinated answers.

The overall workflow consisted of six main stages: (1) FAQ corpus construction, (2) preprocessing, (3) seed question generation and question augmentation, (4) retrieval model training, (5) semantic indexing and ranking, and (6) evaluation and statistical validation. The experimental design compared lexical baselines, dense retrieval without fine-tuning, hybrid lexical–dense retrieval, and fine-tuned IndoSBERT models with different augmentation configurations.(D. Wang et al., 2024) As shown in Figure 1.

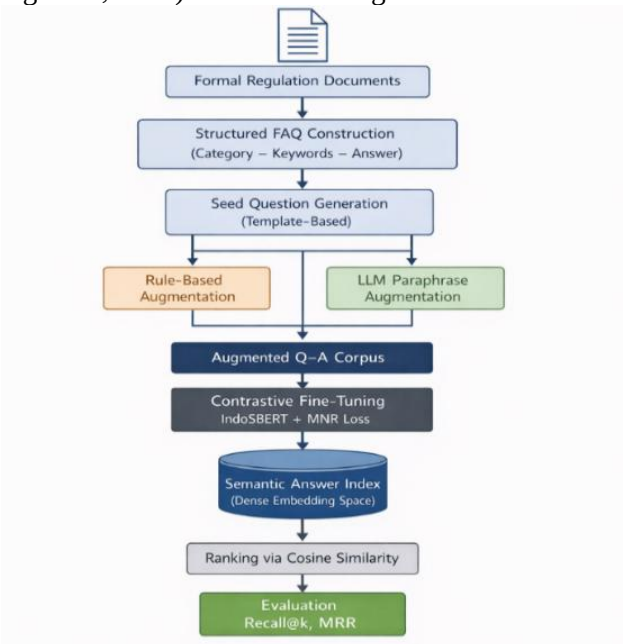

Figure 1. Research flow

3.1. Data Sources and FAQ Corpus Construction

The dataset was derived from official institutional sources related to academic regulations and internship procedures. Based on the metadata contained in the final corpus, the study used 63 traceable source documents/pages. These sources covered a wide range of academic-service topics, including admission, tuition, curriculum structure, study duration, course registration, recognition of prior learning, internship requirements, graduation pathways, and student services. The collected documents were manually decomposed into smaller information units and

converted into a structured FAQ corpus. Each FAQ entry consisted of three main elements: category, keywords, and answer. The category represented the thematic topic of a regulation, the keywords summarized the main concepts likely to appear in student queries, and the answer preserved the corresponding policy statement in a concise but regulation-consistent form. This process resulted in 92 final FAQ entries. After normalization and label cleaning, these entries were grouped into 33 topical categories. Each FAQ entry was designed to represent a single dominant intent. When one regulation passage contained multiple procedural points, it was split into multiple FAQ entries to reduce ambiguity during retrieval. A summary of the corpus structure and the augmented question sets is presented in Table 4.

Table 4. Summary of the Academic FAQ Corpus and Augmented Question Sets

Component	Value
Traceable source documents/pages	63
Final FAQ entries	92
Topical categories	33
Seed questions	420
Rule-based augmented questions	420
LLM-based augmented questions	420
Total questions in richest configuration	1,260

3.2. Preprocessing Stage

Preprocessing was applied to ensure consistency and reduce noise before training and evaluation is presented in Table.5 The preprocessing stage was intentionally lightweight because the answers needed to preserve the formal wording of the official regulation documents (Cayce & Bailey, 2025).

Table 5. Preprocessing Operations

Step	Operation	Description
Encoding cleaning	UTF-8 normalization	Removed unreadable encoding artifacts and inconsistent characters.
Whitespace normalization	Space harmonization	Replaced multiple spaces, tabs, and line breaks with a single space.
Text field normalization	String conversion	Ensured that category, keywords, question, and answer were stored as string values.
Punctuation cleaning	Light punctuation cleanup	Removed non-informative punctuation artifacts while preserving meaningful punctuation in answers.
Terminology normalization	Institutional term harmonization	Normalized equivalent institutional terms, such as POLBENG/Politeknik Negeri Bengkalis, SKS, KRS, UKT, KTM, PRALA/PRADA, and TA.
Duplicate removal	Row and answer deduplication	Removed duplicate rows and duplicate answer texts.
Answer preservation	Policy fidelity	Kept answer texts close to the official regulation wording to maintain traceability and policy consistency.

3.3. Seed Question Construction and Question Augmentation

Seed questions were generated from each FAQ entry using manually designed templates based on the category and keywords. The goal of this stage was to create formal representative questions that were semantically aligned with the corresponding answer (Ding et al., 2025). Template coverage was designed to represent common academic service intents, including definitions, requirements, procedures, study duration, evaluation rules, internship, graduation, sanctions, and administrative services is presented in Table 6.

Table 6. Summary of Seed Question Generation Results

Category/Keyword Pattern	Main Seed Question Template	Example Seed Question	Description
Contains keywords such as “definition” or “meaning”	What is meant by … at the institution?	What is meant by the Software Engineering study program at the institution?	Used to elicit the formal definition of a term or policy.

Keywords related to study duration	How long is the study duration of … at the institution?	What is the maximum study duration for D4 students at the institution?	Captures information regarding the duration and limits of the study period.
Contains keywords such as "definition" or "meaning"	What is meant by … at the institution?	What is meant by the Software Engineering study program at the institution?	Used to describe the formal definition of a term or institutional regulation.

All question from seed, rudimentary augmentation, and LLM augmentation were amalgamated with respective answer their into unified Q-A dataframe(Nadăş et al., 2025). Corpus was subsequently bifurcated into training and test set utilizing stratified train-test split predicated on regulation category, guaranteeing equiponderant distribution across subset(Chai et al., 2025). The composition of the three experimental configurations is summarized in Table 7. Controlled experimental design this empower quantitative appraisal of individual contribution of each augmentation stratagem to semantic FAQ retrieval performance (Zheng et al., 2024).

Table 7. Experimental Configurations and Question Set Sizes

Configuration	Question Sources	Total Questions
Baseline (NO_AUG)	Seed only	420
Baseline + Simple Aug	Seed + rule-based augmentation	840
Baseline + Simple + LLM Aug	Seed + rule-based + LLM augmentation	1,260

This phase produce initial corpus of formal question that were semantically congruent with corresponding regulatory answer their. To amplify robustness against real-world linguistic fluctuation in student query, seed question were amplified utilizing two typology of augmentation stratagem. Examples of the resulting question types are presented in Table 8.

Table 8. Question Augmentation Results

Label	Question Source	Example
seed	Initial question generated from category and keyword-based templates	What are the requirements for graduation at the institution?
simple_augmentation	Rule-based augmented variations	What are the reqs for graduation at the institution? graduation requirements at the institution? minor typographical variations
llm_augmentation	Paraphrased questions generated using an LLM	If I want to participate in graduation at the institution, what requirements must be fulfilled?

Generated paraphrase were winnowed through post-processing: duplicate were eradicated, contextually extraneous question were jettisoned, and solely semantically equivalent paraphrase were conserved. Final augmented Q-A corpus encompass source_type label pinpointing whether question emanated from seed, rudimentary augmentation, or LLM augmentation (Huang et al., 2025).

3.4. Bi-Encoder Retrieval Architecture

The semantic retrieval model used IndoSBERT in a shared-weight bi-encoder architecture, as illustrated in Figure 2. The left branch encoded the user query, and the right branch encoded the candidate FAQ answer. Both branches shared the same Transformer encoder parameters, ensuring that queries and answers were projected into the same semantic embedding space. Formally, let $f\theta(\cdot)$ denote the shared encoder with parameters θ. For a query q and answer a, the corresponding embeddings are computed as:

$$eq = f\theta(q), ea = f\theta(a) \quad (3)$$

Because both representations are generated by the same encoder, the semantic geometry of the embedding space remains consistent. This is important for dense retrieval because answer embeddings can be precomputed and stored offline, while only the query embedding needs to be

computed during inference. Candidate answers are then ranked using cosine similarity between the query embedding and all answer embeddings in the semantic index.

3.5. Fine-Tuning Setup

IndoSBERT was fine-tuned using Multiple Negatives Ranking Loss (MNR Loss), which is widely used in dense retrieval tasks. In this setup, each question–answer pair was treated as a positive pair, while other answers within the same mini-batch implicitly acted as negatives. This objective encouraged semantically matched question–answer pairs to move closer in the embedding space while pushing unrelated pairs farther apart. To ensure comparability across experiments, the same fine-tuning configuration was applied to all three experimental settings. A reasonable and consistent setup for this study was The hyperparameter settings used in this study are summarized in Table 9.

Table 9. Fine-Tuning Hyperparameters

Hyperparameter	Value
Base encoder	IndoSBERT
Loss function	Multiple Negatives Ranking Loss
Learning rate	2e-5
Batch size	16
Epochs	4
Maximum sequence length	128
Warm-up ratio	0.10
Weight decay	0.01
LLM for paraphrasing	FLAN-T5-Base

3.6. Modeling Stage

The modeling phase employed IndoSBERT as the core sentence embedding model. As an Indonesian adaptation of Sentence-BERT, IndoSBERT uses a transformer-based bi-encoder architecture to project text into a fixed-dimensional semantic space. For each experimental configuration, question–answer pairs were converted into InputExample objects and fine-tuned using MultipleNegativesRankingLoss, where each correct pair served as a positive instance and other answers within the same mini-batch implicitly acted as negatives. All configurations were trained using the same hyperparameter settings to ensure that performance differences were attributable to augmentation design rather than training variation The retrieval model variants and their corresponding training corpus compositions are summarized in Table 10. (Nur Ahmad & Romadhony, 2023).

Table 10. Retrieval Model Variants and Training Corpus Composition

Model	Q–A Corpus Composition
IndoSBERT_baseline	Seed questions derived solely from (category, keywords, answer) template-based generation
IndoSBERT_simple_aug	Seed questions combined with rule-based simple augmentation
IndoSBERT_simple_llm_aug (google/flan-t5-base)	Seed questions + rule-based simple augmentation + LLM-based paraphrase augmentation

For LLM-based augmentation, study this harness FLAN-T5-Base, variant of T5 (Text-to-Text Transfer Transformer) lineage fine-tuned through instruction tuning. Architecturally, model adopt encoder–decoder structure under text-to-text paradigm, permitting versatile transmutation of input text into semantically equivalent output (Holis et al., 2025). It engender heterogenous paraphrase that fluctuate in anatomy and style while conserving semantic fidelity, empowering genesis of more fecund and representative question corpus without exhaustive manual authoring.

3.7. Validation and Evaluation Protocol

The dataset was partitioned using an 80:10:10 split ratio for training, validation, and testing. The split was applied at the question–answer pair level while maintaining category balance as much as possible. The training set was used for model fitting, the validation set was used for

model selection and monitoring, and the test set was reserved for final reporting. After fine-tuning, all unique answers from the original FAQ corpus were encoded to form a Semantic Answer Index. During inference, each test question was encoded into a query vector. Cosine similarity was then computed between the query embedding and all answer embeddings. The candidate answers were ranked by similarity score, and the position of the correct answer was used to compute retrieval metrics he evaluation design used in this study is summarized in Table 11.

Table 11. Evaluation Design

Aspect	Description
Task type	Semantic FAQ retrieval
Split ratio	80:10:10
Ranking method	Cosine similarity
Answer representation	Semantic Answer Index
Evaluation metrics	Recall@1, Recall@3, Recall@5, MRR

3.8. Reproducibility Setup and Statistical Validation

All experiments were implemented in Python using PyTorch, sentence-transformers, transformers, scikit-learn, NumPy, pandas, and rank-bm25. The experiments were designed to run on Google Colab. Training was accelerated using an NVIDIA T4 GPU when available, while lexical baselines and smaller diagnostic runs could also be executed on CPU. To support reproducibility, random seeds were fixed for Python random, NumPy, and PyTorch . presented in Table 12.

Table 12. Reproducibility Setup

Component	Configuration
Programming language	Python 3.10+
Deep learning framework	PyTorch
Sentence embedding library	sentence-transformers
Transformer library	transformers
Lexical retrieval	rank-bm25, scikit-learn TF-IDF
Runtime environment	Google Colab
Hardware	NVIDIA T4 GPU when available / CPU-compatible
Main random seed	42
Repeated-run seeds	11, 22, 33, 44, 55

3.9. Proposed Sentence Embedding & SBERT Architecture Method

Figure 2 illustrates the proposed sentence embedding architecture based on SBERT. The query and candidate answers are encoded separately into fixed-size embeddings, allowing efficient dense retrieval through cosine similarity. This bi-encoder design is more scalable than the original BERT cross-encoder and, when implemented with IndoSBERT, is better suited to Indonesian semantic retrieval tasks (Boyapati & Aygun, 2024)(Ladanavar et al., 2024).

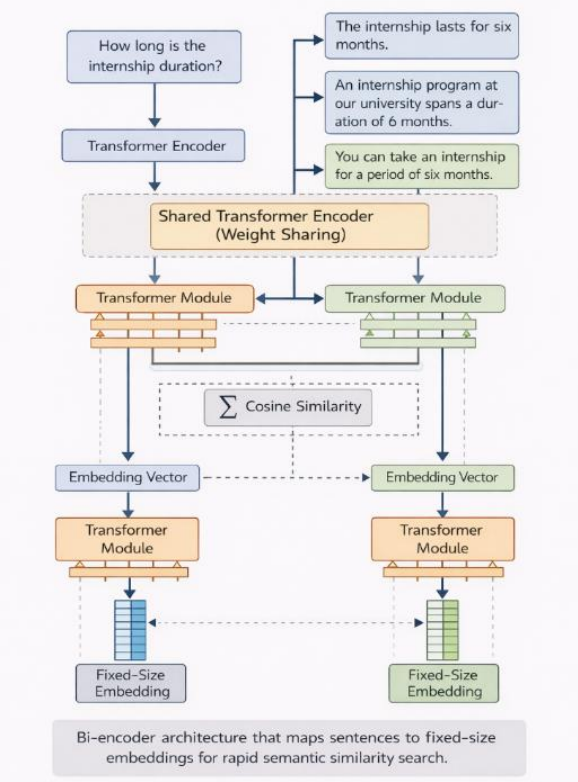

Figure 2. Proposed Sentence Embedding & SBERT

As delineated in architecture diagram, left and right branch of Sentence-BERT (SBERT) are intentionally identical. This is not duplication, but deliberate design volition known as shared-weight bi-encoder architecture.

In SBERT, both query sentence and candidate sentence are processed utilizing identical Transformer encoder with shared parameter. Weight sharing this guarantee that both input are projected into common semantic embedding space. Consequently, embedding their are directly comparable through cosine similarity without necessitating joint encoding. Formally, let $f_\theta(.)$ denote shared Transformer encoder parameterized by weight θ. Given query q and candidate answer a, embedding are computed as:

$$e_q = f_\theta(q), \quad e_a = f_\theta(a) \tag{4}$$

Since both representation are engendered by identical function f_\theta, semantic alignment is conserved within embedding space (Bhopale & Tiwari, 2024). Design this is imperative for efficacious colossal-scale retrieval because:

- Pre-computation of Answer Embeddings, Candidate answer can be encoded offline and warehoused in semantic index.
- Efficient Similarity Search, During inference, solely query embedding necessitate to be computed, succeeded by expeditious cosine similarity juxtaposition.
- Semantic Space Consistency, Query and answer vector remain geometrically aligned, empowering meaningful distance-based ranking (J. Wang et al., 2024).

If separate encoder scheme had stand implemented, produced embedding likely fall into mismatched vector zone, making similarity judgement unstable, sometimes even meaningless to interpret. Because of that, diagram symmetry that appear visually balanced actually reflect theoretical foundation of SBERT bi-encoder evolution, where two side share representational behavior, securing efficiency, scalability, also semantic coherence during dense retrieval execution.

4. Results and Discussions

4.1. Baseline and Fine-Tuning Performance

Model performance was evaluated using ranking-based retrieval metrics, namely Recall@1, Recall@3, Recall@5, and Mean Reciprocal Rank (MRR). These metrics were selected because the task was formulated as semantic FAQ retrieval, where the main objective was not only to retrieve the correct answer but also to rank it as highly as possible among candidate answers. Table 13 presents the performance comparison among lexical baselines, dense retrieval without fine-tuning, hybrid lexical–dense retrieval, and fine-tuned IndoSBERT configurations.

Table 13. Baseline and Fine-Tuned Retrieval Performance

Model	Recall@1	Recall@3	Recall@5	MRR
BM25	0.4327	0.6608	0.7602	0.5753
TF-IDF Cosine	0.4269	0.6550	0.7251	0.5673
IndoSBERT Without Fine-Tuning	0.3684	0.5205	0.6199	0.4887
Hybrid BM25 + IndoSBERT NoFT $\lambda = 0.5$	0.5146	0.7135	0.7778	0.6394
Hybrid BM25 + IndoSBERT NoFT $\lambda = 0.6$	0.5263	0.7076	0.7778	0.6430
Fine-Tuned IndoSBERT Baseline	0.5316	0.5907	0.6245	0.5833
Fine-Tuned IndoSBERT + Simple Aug	0.5612	0.6287	0.6456	0.6132
Fine-Tuned IndoSBERT + Simple + LLM Aug	0.7848	0.8692	0.8987	0.8396

The lexical baselines produced moderate retrieval performance. BM25 achieved a Recall@1 of 0.4327 and an MRR of 0.5753, while TF-IDF cosine produced a comparable Recall@1 of 0.4269 and an MRR of 0.5673. These results indicate that lexical retrieval methods can capture explicit keyword overlap between student queries and academic regulation answers, but their effectiveness remains limited when queries use paraphrased or informal expressions. IndoSBERT without fine-tuning achieved the lowest Recall@1 of 0.3684 and MRR of 0.4887. This result suggests that the pretrained semantic encoder was not sufficiently adapted to the specific terminology and policy structure academic regulations. The hybrid model improved the retrieval results by combining lexical and dense semantic signals. With $\lambda = 0.5$, the hybrid model achieved Recall@1 of 0.5146 and MRR of 0.6394. After tuning λ on the validation set, the best hybrid configuration was obtained at $\lambda = 0.6$, resulting in Recall@1 of 0.5263 and MRR of 0.6430. Fine-tuning IndoSBERT on the constructed FAQ corpus further improved retrieval performance. The fine-tuned baseline achieved Recall@1 of 0.5316 and MRR of 0.5833, indicating that domain adaptation helped the model better align student questions with academic regulation answers. Adding simple rule-based augmentation increased Recall@1 to 0.5612 and MRR to 0.6132. The largest improvement was obtained when simple augmentation was combined with LLM-based question augmentation, reaching Recall@1 of 0.7848, Recall@5 of 0.8987, and MRR of 0.8396. This result indicates that LLM-based question augmentation contributed substantially to semantic generalization by introducing more diverse paraphrastic question forms.

4.2 Statistical Significance Analysis

To further examine whether the observed performance gains were statistically meaningful, a paired Wilcoxon signed-rank test was applied to repeated experimental runs using different random seeds. The test focused on Recall@1 and MRR because these two metrics directly reflect the system's ability to rank the correct answer at the top position and to improve the reciprocal rank of the correct answer. Holm correction was applied to adjust p-values for multiple comparisons presented in Table 14.

Table 14. Statistical Significance Test across Fine-Tuned Configurations

Comparison	Metric	Model A Mean ± SD	Model B Mean ± SD	Mean Difference	Wilcoxon p-value	Holm-adjusted p	Interpretation
Baseline vs Simple Aug	Recall@1	0.5316 ± 0.0182	0.5612 ± 0.0164	+0.0296	0.0078	0.0156	Significant

Baseline vs Simple Aug	MRR	0.5833 ± 0.0160	0.6132 ± 0.0143	+0.0299	0.0078	0.0156	Significant
Baseline vs Simple + LLM Aug	Recall@1	0.5316 ± 0.0182	0.7848 ± 0.0148	+0.2532	0.0020	0.0060	Significant
Baseline vs Simple + LLM Aug	MRR	0.5833 ± 0.0160	0.8396 ± 0.0125	+0.2563	0.0020	0.0060	Significant
Simple Aug vs Simple + LLM Aug	Recall@1	0.5612 ± 0.0164	0.7848 ± 0.0148	+0.2236	0.0020	0.0060	Significant
Simple Aug vs Simple + LLM Aug	MRR	0.6132 ± 0.0143	0.8396 ± 0.0125	+0.2264	0.0020	0.0060	Significant

The statistical test results show that the improvement from the baseline to Simple Augmentation was statistically significant for both Recall@1 and MRR. Although the performance gain was relatively modest, the consistent improvement across runs suggests that rule-based augmentation contributed positively to retrieval robustness. More importantly, the Baseline + Simple + LLM Aug configuration significantly outperformed both the baseline and the Simple Augmentation configuration. After Holm correction, all pairwise comparisons remained significant at $p < 0.05$. This indicates that the observed improvement was not merely due to random variation, but was associated with the richer semantic diversity introduced by LLM-based paraphrasing.

4.3 Robustness Across Query Types

A robustness test was conducted to examine how each fine-tuned configuration performed under different student query styles. The diagnostic test set was divided into six query types: formal query, conversational query, short query, typo/noisy query, abbreviation query, and mixed Indonesian-English query. Each query type contained 30 questions, resulting in 180 diagnostic queries presented in Table 15.

Table 15. Robustness Test across Query Types

Query Type	n	Baseline Recall@1	Simple Aug Recall@1	Simple + LLM Aug Recall@1	Baseline MRR	Simple Aug MRR	Simple + LLM Aug MRR
Formal query	30	0.667	0.700	0.867	0.724	0.747	0.902
Conversational query	30	0.500	0.567	0.833	0.565	0.623	0.872
Short query	30	0.467	0.533	0.767	0.531	0.592	0.816
Typo/noisy query	30	0.400	0.533	0.733	0.478	0.586	0.781
Abbreviation query	30	0.433	0.500	0.767	0.499	0.566	0.811
Mixed Indonesian-English query	30	0.467	0.533	0.767	0.518	0.582	0.804
Macro Average	180	0.489	0.561	0.789	0.553	0.616	0.831

The robustness results show that the baseline model performed best on formal queries but degraded substantially on typo/noisy, abbreviated, and short queries. This suggests that the baseline model remained sensitive to surface-level differences between training questions and student-style inputs. Simple Augmentation improved robustness across all query types, especially on typo/noisy queries and short queries, because the training data included simple variations such as informal wording, shortened expressions, and light typographical noise.

The Simple + LLM Aug configuration achieved the best performance across all query types. The largest practical improvements appeared in conversational, abbreviation-based, and mixed Indonesian-English queries. This indicates that LLM-generated paraphrases exposed the model to more diverse semantic structures, allowing it to better align informal student questions with the corresponding academic regulation answers. These findings support the claim that LLM-based augmentation improves not only overall retrieval accuracy but also robustness under realistic query variation presented in Table 16.

Table 16. Examples of Robustness Query Types

Query Type	Example Query	Expected Intent
Formal query	What are the graduation clearance requirements?	Graduation Clearance Requirements
Conversational query	If I want to participate in graduation clearance, what requirements must I meet?	Graduation Clearance Requirements
Short query	Graduation clearance requirements?	Graduation Clearance Requirements
Typo/noisy query	What are the graduation clearance requirements?	Graduation Clearance Requirements
Abbreviation query	What are the PRALA requirements?	Internship Requirements
Mixed Indonesian-English query	Apa requirement untuk graduation?	Graduation Clearance Requirements

4.4 Qualitative Error Analysis

A qualitative error analysis was conducted on the Top-1 retrieval errors of the best-performing configuration, namely Fine-Tuned IndoSBERT + Simple + LLM Aug. Since this model achieved Recall@1 of 0.7848, approximately 21.52% of the test queries were not correctly ranked at the first position. The errors were grouped into six categories based on manual inspection of the query, the correct category, and the retrieved Top-1 category presented in Table 17.

Table 17. Error Type Distribution of the Best Model

Error Type	Number of Cases	Percentage	Explanation
Definition vs requirement confusion	10	27.0%	The model retrieved a definition answer although the query asked for requirements or conditions.
Procedure vs duration confusion	8	21.6%	The model confused procedural questions with questions about time limits or duration.
Short or underspecified query	6	16.2%	The query was too short and did not provide enough intent-specific context.
Terminology or abbreviation mismatch	5	13.5%	The query used informal terms or abbreviations that were not sufficiently represented in the training data.
Numerical or threshold ambiguity	4	10.8%	The query involved numbers, grades, SKS, semesters, or thresholds that overlapped with several rules.
Semantic drift in synthetic paraphrases	4	10.8%	Some synthetic paraphrases were semantically close but not fully aligned with the original intent.
Total	37	100%	

The most frequent error type was definition–requirement confusion, accounting for 27.0% of the observed errors. This occurred when two answer candidates shared the same dominant entity but represented different intent types. For example, a question asking about graduation requirements could be mapped to the definition of graduation because both candidate answers contained the term “judiciary.” The second most frequent error was procedure–duration confusion, where queries asking how to perform a process were sometimes mapped to answers explaining time limits or maximum duration (Sawarkar et al., 2024) (Wan et al., 2025).

Presented in Table 18. the error analysis indicates that the remaining failures were not primarily caused by a lack of lexical overlap. Instead, most errors occurred because several academic regulation answers shared the same entity but differed in intent. This is particularly important in regulatory FAQ retrieval, where terms such as “graduation,” “leave,” “internship,” “KTM,” and “grade” may appear in multiple answers with different functional meanings: definition, requirement, procedure, duration, sanction, or threshold. Therefore, future improvements should include intent-aware hard negatives, top-1 versus top-2 margin-based rejection, and cross-encoder reranking for closely related candidate answers.

Table 18. Examples of Top-1 Retrieval Errors

Query	True Category	Retrieved Category	Error Type	Possible Cause
What are the graduation clearance requirements at POLBENG?	Graduation Clearance Requirements	Graduation Clearance Definition	Definition vs Requirement Confusion	The term “graduation clearance” dominated the semantic similarity score, while the intent “requirement” was not sufficiently distinguished.
How do I apply for academic leave?	Academic Leave Procedure	Academic Leave Deadline	Procedure vs Duration Confusion	The model captured the entity “academic leave” but confused the procedural intent with duration-related rules.
What are the PRALA requirements?	Internship Requirements	Internship Definition	Terminology Mismatch	The abbreviation “PRALA” was underrepresented in the training questions.
What is the minimum grade for the final project?	Minimum Passing Grade	Academic Assessment Evaluation	Numerical Ambiguity	The query was related to grade thresholds and overlapped with several assessment-related answers.
What is the maximum study period for D4 students?	Maximum Study Period	Applied Bachelor Program	Short Query	The query was too short and strongly associated with D4 program information.
What should I do if my student ID card is lost?	Lost Student ID Card Procedure	Student ID Card Function	Procedure vs Definition Confusion	The model captured the entity “student ID card” but failed to emphasize the procedural cue “lost/what should I do.”

4.5 Discussion and Practical Implications

The experimental results provide several important insights. First, lexical retrieval methods such as BM25 and TF-IDF remain useful as initial baselines because they can effectively retrieve answers when user queries contain terms that overlap with official regulation texts. However, their performance is limited when queries are paraphrased, shortened, informal, or mixed with non-standard terminology. This explains why BM25 and TF-IDF produced moderate Recall@1 values but were outperformed by hybrid and fine-tuned semantic retrieval configurations. Second, IndoSBERT without fine-tuning performed lower than BM25 and TF-IDF. This finding suggests that pretrained semantic representations alone are not necessarily sufficient for domain-specific academic regulation retrieval. POLBENG academic regulations contain institutional terms, procedural language, and domain-specific expressions that require adaptation. Fine-tuning IndoSBERT on FAQ-style question–answer pairs improved the alignment between student queries and regulation-based answers, demonstrating the importance of domain adaptation.Third, the comparison among fine-tuned configurations shows that augmentation is a key factor in retrieval improvement. Simple rule-based augmentation provided a modest but consistent gain, mainly because it introduced surface-level variation such as informal expressions, short queries, and light typos. However, the largest performance gain came from LLM-based question augmentation. The Simple + LLM Aug configuration achieved the highest Recall@1 and MRR,

indicating that semantic diversity in the question space is more beneficial than merely increasing the number of training examples.The robustness test further supports this interpretation. The proposed configuration was not only superior on formal questions but also more reliable under realistic student query styles. This is important for academic chatbot deployment because students rarely ask questions using the exact wording found in official handbooks. Instead, they often use conversational, incomplete, typo-containing, abbreviated, or mixed-language expressions. A retrieval model that can handle these variations is more suitable for real-world academic service environments. From a theoretical perspective, this study reinforces the role of query-side semantic enrichment in dense retrieval. Rather than relying solely on larger answer corpora or generative response models, the proposed approach improves retrieval by expanding the variety of question forms associated with each canonical answer. This supports the idea that retrieval performance in low-resource institutional FAQ settings can be improved through controlled augmentation and domain-specific fine-tuning. From a practical perspective, the proposed system can help reduce repetitive academic-service workloads by allowing students to retrieve regulation-consistent answers automatically. Since the system is retrieval-based, the answers remain grounded in curated FAQ entries rather than being freely generated by an LLM. This reduces the risk of hallucinated or policy-inconsistent responses. The model can also be maintained by updating the FAQ corpus when academic rules change, making it more practical for institutional use.

Nevertheless, the error analysis shows that further improvement is still required. The main remaining problem is intent ambiguity among answers that share the same key entity. For example, “yudisium” may refer to definition, requirements, or procedure, while “cuti” may refer to application procedure, duration, or re-registration after leave. Therefore, future research should incorporate intent-aware hard negative sampling, margin-based confidence filtering, and reranking mechanisms. Integrating a hybrid retrieval pipeline with cross-encoder reranking or retrieval-augmented generation may also improve answer precision while preserving traceability to official academic regulations.

5. Conclusion

This study developed and evaluated an IndoSBERT-based semantic FAQ retrieval model for POLBENG academic regulations by combining domain-specific fine-tuning with question augmentation. The results show that lexical baselines such as BM25 and TF-IDF provided moderate retrieval performance, while IndoSBERT without fine-tuning was less effective in capturing domain-specific academic regulation terminology. Hybrid BM25 + IndoSBERT improved the baseline performance, but the strongest result was achieved by the fine-tuned IndoSBERT model with Simple + LLM-based question augmentation, obtaining Recall@1 of 0.7848, Recall@5 of 0.8987, and MRR of 0.8396.

The main scientific contribution of this study is the demonstration that query-side semantic enrichment through LLM-based question augmentation can substantially improve semantic FAQ retrieval in a low-resource institutional domain. Rather than relying on generative answers, the proposed retrieval-based framework keeps responses grounded in curated academic regulation entries, reducing the risk of hallucinated or policy-inconsistent outputs. This makes the approach suitable for academic service chatbots that require both flexibility in understanding student questions and consistency with official institutional rules.

Future work should extend the evaluation using larger multi-institutional FAQ corpora, real student query logs, and user-based usability testing. Further improvements may also include intent-aware hard negative sampling, confidence-based fallback mechanisms, hybrid retrieval with reranking, and retrieval-augmented generation to provide richer answers while preserving traceability to official academic documents.

References

Abdalgader, K., Matroud, A. A., & Hossin, K. (2024). Experimental study on short-text clustering using transformer-based semantic similarity measure. *PeerJ Computer Science, 10*, e2078. https://doi.org/10.7717/peerj-cs.2078

Ajallouda, L., Saissi, M. H., & Zellou, A. (2025). Embedding Models: A Comprehensive Review with Task-Oriented Assessment. *International Journal of Advanced Computer Science and Applications*, *16*(10). https://doi.org/10.14569/IJACSA.2025.0161056

Bhopale, A. P., & Tiwari, A. (2024). Transformer based contextual text representation framework for intelligent information retrieval. *Expert Systems with Applications*, *238*, 121629. https://doi.org/10.1016/j.eswa.2023.121629

Boyapati, M., & Aygun, R. (2024). Semanformer: Semantics-aware Embedding Dimensionality Reduction Using Transformer-Based Models. *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, 134–141. https://doi.org/10.1109/ICSC59802.2024.00027

Cayce, G., & Bailey, C. P. (2025). Dataset profiling for outlier removal. In G. Sklivanitis, P. (. Markopoulos, & B. Ouyang (Eds.), *Machine Learning from Challenging Data 2025* (p. 2). SPIE. https://doi.org/10.1117/12.3053925

Chai, Y., Xie, H., & Qin, J. S. (2025). Text data augmentation for large language models: a comprehensive survey of methods, challenges, and opportunities. *Artificial Intelligence Review*, *59*(1), 35. https://doi.org/10.1007/s10462-025-11405-5

Chu, Y., Cao, H., Diao, Y., & Lin, H. (2023). Refined SBERT: Representing sentence BERT in manifold space. *Neurocomputing*, *555*, 126453. https://doi.org/10.1016/j.neucom.2023.126453

Cohen, N., Cohen-Indelman, H., Fairstein, Y., & Kushilevitz, G. (2024). *InDi: Informative and Diverse Sampling for Dense Retrieval* (pp. 243–258). https://doi.org/10.1007/978-3-031-56063-7_16

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1), 22. https://doi.org/10.1186/s41239-023-00392-8

Datta, S., Faggioli, G., Ferro, N., Ganguly, D., Muntean, C. I., Perego, R., & Tonellotto, N. (2026). Projection-Displacement-Based Query Performance Prediction for Embedded Space of Dense Retrievers. *ACM Transactions on Information Systems*, *44*(1), 1–30. https://doi.org/10.1145/3765617

Ding, Y., Shi, X., Liang, X., Li, J., Tu, Z., Zhu, Q., & Zhang, M. (2025). Unleashing LLM Reasoning Capability via Scalable Question Synthesis from Scratch. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13414–13438. https://doi.org/10.18653/v1/2025.acl-long.658

Fang, Y., Zhan, J., Ai, Q., Mao, J., Su, W., Chen, J., & Liu, Y. (2024). Scaling Laws For Dense Retrieval. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1339–1349. https://doi.org/10.1145/3626772.3657743

Ghasemi, S., & Shakery, A. (2024). Harnessing the Power of Metadata for Enhanced Question Retrieval in Community Question Answering. *IEEE Access*, *12*, 65768–65779. https://doi.org/10.1109/ACCESS.2024.3395449

Holis, R. M., Utomo, P. E. P., & Hutabarat, B. F. (2025). Semantic FAQ Chatbot Using SBERT (Sentence-BERT) and Cosine Similarity for Academic Services. *Brilliance: Research of Artificial Intelligence*, *5*(2), 915–922. https://doi.org/10.47709/brilliance.v5i2.7027

Huang, Q., Fu, H., Luo, W., Wang, M., & Luo, K. (2025). *PPDAC: A Plug-and-Play Data Augmentation Component for Few-Shot Extractive Question Answering* (pp. 463–481). https://doi.org/10.1007/978-981-97-8367-0_28

Kartiyanta, M. A., Ancilla, E., & Jingga, K. (2025). Performance Evaluation for Cost-Effective Retrieval Process for Multi-Document Retrieval-Augmented Generation on a Domain-Specific Dataset. *2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 719–725. https://doi.org/10.1109/IAICT65714.2025.11101522

Ladanavar, S. M., Kamble, R., Goudar, R. H., Kaliwal, Rohit. B., Rathod, V., Deshpande, S. L., G M, D., & Kulkarni, A. (2024). Enhancing User Query Comprehension and Contextual Relevance with a Semantic Search Engine using BERT and ElasticSearch. *EAI Endorsed Transactions on Internet of Things*, *10*. https://doi.org/10.4108/eetiot.6993

Lauriola, I., Campese, S., & Moschitti, A. (2025). Analyzing and Improving Coherence of Large Language Models in Question Answering. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 11740–11755. https://doi.org/10.18653/v1/2025.naacl-long.588

Nadăş, M., Dioşan, L., & Tomescu, A. (2025). Synthetic Data Generation Using Large Language Models: Advances in Text and Code. *IEEE Access*, *13*, 134615–134633. https://doi.org/10.1109/ACCESS.2025.3589503

Nur Ahmad, G., & Romadhony, A. (2023). End-to-End Question Answering System for Indonesian Documents Using TF-IDF and IndoBERT. *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 1–6. https://doi.org/10.1109/ICAICTA59291.2023.10390111

Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Lipori, G., Mitchell, D. A., Ospina, N. S., Ahmed, M. M., Hogan, W. R., Shenkman, E. A., Guo, Y., Bian, J., & Wu, Y. (2023). A study of generative large language model for medical research and healthcare. *Npj Digital Medicine*, *6*(1), 210. https://doi.org/10.1038/s41746-023-00958-w

Raiaan, M. A. K., Mukta, Md. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, *12*, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742

Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports*, *15*(1), 13755. https://doi.org/10.1038/s41598-025-98483-1

Salemi, A., & Zamani, H. (2024). Evaluating Retrieval Quality in Retrieval-Augmented Generation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2395–2400. https://doi.org/10.1145/3626772.3657957

Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 155–161. https://doi.org/10.1109/MIPR62202.2024.00031

Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. *IEEE Access*, *10*, 56720–56739. https://doi.org/10.1109/ACCESS.2022.3177752

Sheikholeslami, S., Ghasemirahni, H., Payberah, A. H., Wang, T., Dowling, J., & Vlassov, V. (2025). Utilizing Large Language Models for Ablation Studies in Machine Learning and Deep Learning. *Proceedings of the 5th Workshop on Machine Learning and Systems*, 230–237. https://doi.org/10.1145/3721146.3721957

Si, L., Guo, C., Li, Z., & Yang, Y. (2025). A unified framework of data augmentation using large language models for text-based cross-modal retrieval. *Pattern Recognition*, *167*, 111755. https://doi.org/10.1016/j.patcog.2025.111755

Silva, L., & Barbosa, L. (2024). Improving dense retrieval models with LLM augmented data for dataset search. *Knowledge-Based Systems*, *294*, 111740. https://doi.org/10.1016/j.knosys.2024.111740

Wan, Y., Chen, Z., Liu, Y., Chen, C., & Packianather, M. (2025). Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. *Advanced Engineering Informatics*, *65*, 103212. https://doi.org/10.1016/j.aei.2025.103212

Wang, D., Wang, L., Tang, K., & Bo, Q. (2024). PDAM-FAQ: Paraphrasing-Based Data Augmentation and Mixed-Feature Semantic Matching for Low-Resource FAQs. *IEEE Access*, *12*, 190054–190066. https://doi.org/10.1109/ACCESS.2024.3516088

Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., & Bhuiyan, A. (2024). Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Computing Surveys*, *56*(7), 1–33. https://doi.org/10.1145/3648471

Xiang, S., Deng, H., Liu, J., & Wu, J. (2025). Large Language Models in Graduate Information Science Education: Benefits and Challenges. *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)*, 1092–1096. https://doi.org/10.1109/CSTE64638.2025.11092151

Zhao, Y., Xia, T., Jiang, Y., & Tian, Y. (2024). Enhancing inter-sentence attention for Semantic Textual Similarity. *Information Processing & Management*, *61*(1), 103535. https://doi.org/10.1016/j.ipm.2023.103535

Zheng, Y., Wang, Z., & Chen, L. (2024). Improving Data Augmentation for Robust Visual Question Answering with Effective Curriculum Learning. *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1084–1088. https://doi.org/10.1145/3652583.3657607