



Development of a Question Answering System Using the Best Matching 25 (BM25) Method to Support Final Project Completion for Informatics Engineering Students at UNISSULA

Iqbal Nur Al Sandi^{1*}, Ndaru Senopati Nusa², Muhammad Farrel Ardan³, Bagus Satrio Waluyo Poetro⁴, Moch Taufik⁵

¹⁻⁵ Universitas Islam Sultan Agung, Indonesia
email: senopatiip20@gmail.com¹

Article Info :

Received:
21-08-2026
Revised:
03-09-2026
Accepted:
16-09-2026

Abstract

This study developed and evaluated a BM25-based question-answering system to support final-project information retrieval for Informatics Engineering students at Universitas Islam Sultan Agung. The system used the official Final Project Guidebook and Supervision Logbook as its knowledge corpus and integrated document reconstruction, heading-aware chunking, paragraph overlap, lexical preprocessing, synonym-based query expansion, heading-weighted indexing, BM25 ranking, and source-grounded answer presentation. Evaluation was conducted using 50 manually mapped ground-truth queries and measured through Recall@1, Recall@3, Recall@5, and Mean Reciprocal Rank. The results showed that 41 queries retrieved the relevant passage at rank 1, eight at rank 2, and one at rank 3. Recall@1 reached 82.00%, while Recall@3 and Recall@5 both achieved 100.00%, with an MRR of 0.9067. These findings indicate that the proposed lexical retrieval architecture can reliably recover authoritative academic information while maintaining transparency, interpretability, and source traceability. The remaining limitation concerns first-rank discrimination among procedurally similar passages, suggesting that future improvements should prioritize reranking or hybrid retrieval without sacrificing complete Top-3 coverage. The approach also provides a reproducible baseline for evaluating more complex semantic retrieval methods in academic information environments.

Keywords: Academic Information, BM25, Information Retrieval, Question Answering System, Text Processing.



©2026 Authors.. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.
(<https://creativecommons.org/licenses/by-nc/4.0/>)

INTRODUCTION

The rapid digitalization of higher education has transformed the way universities organize, distribute, and provide access to academic information. Institutional regulations, academic guidelines, supervision records, and administrative procedures are increasingly stored in digital repositories, yet the availability of digital documents does not automatically ensure efficient information access. Students may still encounter difficulties locating specific procedural requirements because relevant information is often distributed across lengthy and structurally complex documents, while conventional document-search mechanisms rely heavily on manual navigation or literal keyword matching (Atmojo & Kunang, 2023). This condition creates a need for intelligent academic information systems capable of identifying relevant passages accurately and presenting them in a form that is directly accessible to users.

Question-answering systems offer a promising approach to this problem because they enable users to express information needs through natural-language questions rather than manually examining entire documents. Their effectiveness, however, depends substantially on the ability of the retrieval component to locate authoritative passages that correspond closely to the intended meaning of a query. Information relevance cannot be determined solely by the presence of identical terms because a document becomes useful only when its content adequately satisfies a particular information need (Buckland & de Fremery, 2024). Academic procedural information intensifies this challenge because terms related to registration, supervision, seminars, examinations, and graduation requirements may repeatedly appear in different sections while referring to distinct administrative conditions.

Natural Language Processing (NLP) and Information Retrieval (IR) provide computational mechanisms for transforming natural-language queries and document collections into representations that can be processed, compared, and ranked. One established lexical ranking method is Best Matching 25 (BM25), which estimates document relevance through the interaction of term frequency, inverse document frequency, and document-length normalization (Darmawan et al., 2025). BM25 remains relevant in contemporary retrieval architectures because its ranking mechanism is computationally efficient, interpretable, and capable of generating strong lexical candidates even when more complex neural approaches are available (Sager et al., 2025). Its performance nevertheless depends on the quality of document representation and query processing, meaning that ranking effectiveness cannot be attributed to the retrieval algorithm alone.

Document preparation represents an important stage because institutional academic documents frequently contain hierarchical headings, tables, images, and extended procedural explanations that may complicate text extraction and segmentation. Inappropriate chunking may separate a requirement from its explanatory context or combine unrelated procedural information into a single retrieval unit, thereby reducing the precision of subsequent ranking. Research on retrieval-oriented architectures has demonstrated that structural segmentation and chunking strategies can materially influence whether relevant information remains available as a coherent retrievable passage (Agustiani et al., 2026). Preserving document structure through heading-aware segmentation is therefore particularly important for academic guidebooks whose hierarchical sections distinguish conceptually similar but procedurally different information.

Lexical variation between user questions and institutional documents creates an additional challenge because students do not always employ the formal terminology used in official academic regulations. Text normalization, stopword removal, token filtering, and other preprocessing procedures can influence the lexical features retained for subsequent retrieval, especially in Indonesian-language document collections (Setiawan et al., 2025). Query expansion provides another mechanism for reducing vocabulary mismatch by enriching a user's original query with synonymous or semantically associated terms without altering the lexical composition of the indexed corpus (Suparto et al., 2025). Recent research has likewise shown that expanded queries can provide additional retrieval evidence when the original information request is short, sparse, or expressed differently from the terminology appearing in source documents (Wang et al., 2023).

Recent developments in information retrieval increasingly employ dense representations, hybrid retrieval, and reranking architectures to capture semantic relations that may not be represented through direct lexical overlap. Dense passage retrieval demonstrates the potential of learned representations for retrieving semantically related passages in large-scale question-answering environments, although such architectures introduce additional training, computational, and implementation requirements (Karpukhin et al., 2020). Hybrid approaches continue to retain lexical retrieval because BM25 can provide effective candidate generation before more computationally intensive ranking stages are applied (Sager et al., 2025; Schofield et al., 2025). These developments indicate that the appropriateness of a retrieval method should be assessed according to corpus characteristics, information needs, interpretability requirements, and measurable ranking performance rather than assuming that greater model complexity necessarily produces better institutional retrieval.

A research gap remains in the evaluation of lightweight and interpretable lexical question-answering systems for narrowly bounded Indonesian academic corpora in which answers must be traceable directly to official institutional documents. Retrieval evaluation requires explicit relevance judgments because plausible textual similarity cannot establish whether a system consistently places the intended passage at a sufficiently high ranking position (Gienapp et al., 2024). Ranking-sensitive metrics are particularly important because a system may successfully retrieve the correct passage but still reduce practical usability when that passage appears below stronger lexical competitors, making measures such as Recall at K and Mean Reciprocal Rank appropriate for evaluating retrieval quality (Moffat, 2022). This issue is especially relevant to the Informatics Engineering Program at Universitas Islam Sultan Agung (UNISSULA), where students require accurate access to procedural information contained in the official Final Project Guidebook and Supervision Logbook.

This study therefore develops and evaluates a BM25-based question-answering system designed to support final-project completion among Informatics Engineering students at UNISSULA using official institutional documents as the exclusive knowledge corpus. The proposed architecture

integrates document reconstruction, heading-aware chunking, paragraph overlap, lexical preprocessing, synonym-based query expansion, heading-weighted indexing, BM25 ranking, and source-grounded answer presentation within a single reproducible retrieval pipeline. Its novelty lies in examining whether a transparent lexical architecture can achieve reliable passage-level retrieval within a small Indonesian academic corpus while preserving structural context, institutional source fidelity, and direct traceability of the retrieved information. System effectiveness is quantitatively evaluated using 50 manually mapped ground-truth queries through Recall@1, Recall@3, Recall@5, and Mean Reciprocal Rank, providing an empirical basis for assessing both retrieval coverage and the ranking position of the first relevant passage.

RESEARCH METHODS

This study employed an empirical system-development approach combined with quantitative information-retrieval evaluation to develop and assess a BM25-based question-answering system for final-project academic information. The research corpus consisted exclusively of the official Final Project Guidebook and Supervision Logbook of the Informatics Engineering Program at Universitas Islam Sultan Agung (UNISSULA), which were reconstructed from PDF into DOCX format while preserving their original content and structural sequence. Document preparation involved heading-aware chunking based on Heading 1, Heading 2, and Heading 3, with a maximum chunk length of 500 characters, a minimum threshold of 60 characters, and paragraph overlap between adjacent chunks to preserve contextual continuity; structural segmentation is important for maintaining coherent information units in retrieval systems (Agustiani et al., 2026; Fahreza & Anwar, 2026). Text preprocessing consisted of Unicode NFKC normalization, case folding, regular-expression-based cleaning, tokenization, Indonesian stopword removal, and filtering of tokens shorter than two characters, while each resulting chunk retained its textual content, heading path, and unique identifier to ensure source traceability. Query-specific synonym expansion was subsequently applied to reduce lexical mismatch between informal student expressions and terminology contained in official institutional documents, whereas the indexed corpus remained unexpanded to preserve its original lexical distribution (Setiawan et al., 2025; Suparto et al., 2025).

The retrieval process was implemented using BM25Okapi from the Rank-BM25 library, with the parameters $k1 = 1.0$ and $b = 0.3$, while heading tokens received additional structural weighting to strengthen topical relevance during ranking. Each user query underwent the same preprocessing procedure applied to the corpus before synonym-based expansion, after which BM25 relevance scores were calculated for all indexed chunks, candidates with scores less than or equal to zero were excluded, and the five highest-scoring passages were retained as the final ranked results. Sequential chunks originating from the same heading could be reconstructed when necessary, while duplicated overlapping text was removed to maintain readability and fidelity to the original source. System performance was evaluated using 50 representative queries manually mapped to ground-truth document or passage identifiers derived from the official Final Project Guidebook and Supervision Logbook. Retrieval effectiveness was measured using Recall@1, Recall@3, Recall@5, and Mean Reciprocal Rank (MRR), where Recall@K identified whether the relevant passage appeared within the top-K results and MRR measured the reciprocal position of the first relevant passage, thereby providing both coverage- and ranking-sensitive evaluation of system performance (Moffat, 2022).

RESULTS AND DISCUSSION

Document Preparation, Structural Chunking, and Text Preprocessing

The first stage of system development produced a structured document corpus from the two official sources used in this study, namely the Final Project Guidebook and the Supervision Logbook of the Informatics Engineering Program at Universitas Islam Sultan Agung. Both documents were initially available in PDF format and contained combinations of textual content, tables, images, headings, and page-layout elements that could interfere with consistent text extraction. The documents were consequently reconstructed into DOCX format while preserving their informational content and original sequence rather than rewriting or summarizing the source material. This conversion established a more stable textual representation for subsequent segmentation, preprocessing, and indexing, which is important because institutional information systems require accessible source structures before effective retrieval can occur (Atmojo & Kunang, 2023). Similar

academic information applications also indicate that the reliability of an interactive information service is closely associated with the quality of the institutional knowledge source supplied to the system (Mursidah et al., 2022; Bunda et al., 2025).

The comparison between the initial PDF representation and the converted DOCX dataset demonstrates that document preparation was not merely a technical file-format change but a mechanism for controlling the quality of information entering the retrieval pipeline. Complex PDF layouts can disrupt reading order or separate related textual components during automated extraction, creating errors that cannot subsequently be corrected by the ranking algorithm when the relevant textual evidence has already been lost or fragmented. The converted representation maintained the original procedural information while providing clearer structural boundaries that could be recognized programmatically during chunk formation. This approach is particularly relevant for document-oriented question-answering because retrieval performance depends on the accessibility and traceability of the information contained in the source collection (Lawami et al., 2024; Bunda et al., 2025). The resulting DOCX corpus consequently functioned as a controlled intermediate representation between the original institutional documents and the computational retrieval architecture.

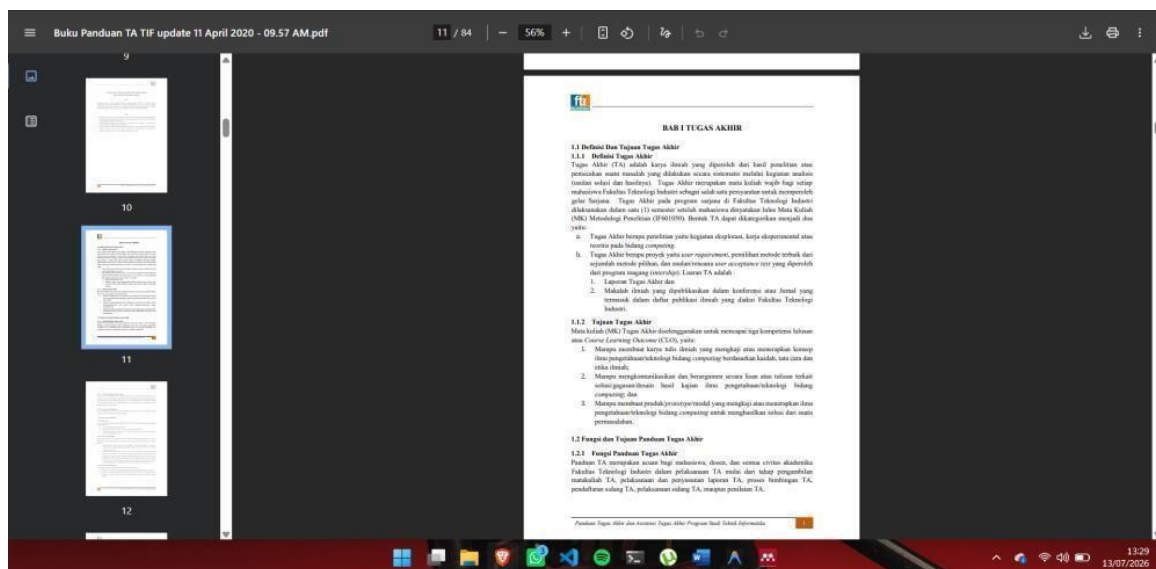


Figure 2. Initial PDF Dataset

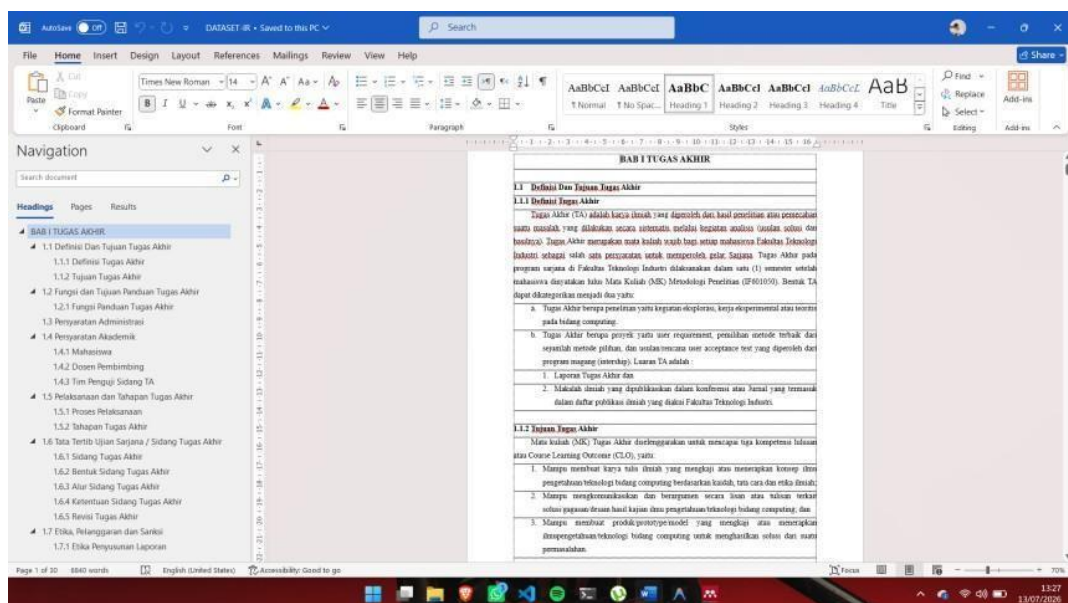


Figure 3. Converted DOCX Dataset

Structural segmentation was subsequently performed using a heading-aware chunking strategy in which Heading 1, Heading 2, and Heading 3 acted as boundaries for identifying changes in document context. Each chunk was limited to a maximum length of 500 characters, while chunks containing fewer than 60 characters were excluded to prevent isolated headings or short transitional fragments from becoming weak retrieval units. When a segment exceeded the specified limit, the final paragraph of the preceding chunk was retained at the beginning of the subsequent chunk to create contextual overlap. The overlap mechanism was intended to reduce the possibility that a procedural requirement would become separated from the condition, qualification, or explanation immediately surrounding it. This treatment is consistent with evidence that chunking strategy influences whether relevant evidence remains coherent and retrievable in retrieval-oriented architectures (Agustiani et al., 2026), while structural approaches can provide advantages when technical or academic documents contain hierarchical sections with heterogeneous information (Fahreza & Anwar, 2026).

The chunking process also preserved a heading path and unique identifier for each retained textual unit, enabling retrieved passages to remain associated with their original structural location. Such metadata are important in an academic procedural corpus because similar expressions may occur under different sections while conveying different administrative meanings, such as requirements for registration, supervision, proposal seminars, or examinations. A lexical match that ignores structural location may consequently rank a passage containing highly similar vocabulary even when it belongs to an inappropriate procedural section. Structural metadata create an additional layer of interpretability because retrieval errors can be examined in relation to both their textual content and their location in the source hierarchy. The approach also supports efficient indexing because structured document collections can be transformed into reusable retrieval units instead of requiring the complete source documents to be scanned for every user request (Mallia et al., 2022).



```
level = get_heading_level(style)

if level is not None:
    flush(buffer, " > ".join(active_headings) if active_headings else "Umum")
    buffer = []

    ...

    active_headings = active_headings[:level - 1]
    while len(active_headings) < level - 1:
        active_headings.append("?")

    active_headings.append(numbered_heading)
```

Figure 4. Heading-Aware Chunking Implementation

The next processing stage standardized the lexical representation of both indexed passages and incoming questions through a deterministic preprocessing procedure. The implemented sequence applied Unicode NFKC normalization, case folding, removal of punctuation and non-alphanumeric characters, tokenization, Indonesian stopwords removal, and filtering of tokens shorter than two characters. These operations reduced superficial textual variation while retaining words expected to contribute meaningful evidence during BM25 scoring, an important consideration because preprocessing choices can materially alter the features available to subsequent Indonesian text-processing algorithms (Setiawan et al., 2025). Stopword removal was particularly relevant because highly frequent functional terms generally provide limited discriminative value for identifying the procedural section sought by a user, although the selection of stopwords must remain sensitive to characteristics of the target corpus (Solihin et al., 2026). The complete transformation is illustrated by

the original query “*Bagaimana syarat mendaftar Tugas Akhir?*”, whose successive representations are retained in Table 1.

Table 1. Example of Text Preprocessing Results

Stage	Result
Original Text	Bagaimana syarat mendaftar Tugas Akhir?
Case Folding	bagaimana syarat mendaftar tugas akhir
Cleaning	bagaimana syarat mendaftar tugas akhir
Tokenization	["bagaimana", "syarat", "mendaftar", "tugas", "akhir"]
Stopwords Removal	["syarat", "mendaftar", "tugas", "akhir"]
Query Expansion	["syarat", "mendaftar", "pendaftaran", "daftar", "registrasi", "tugas", "akhir"]

Table 1 demonstrates how the preprocessing pipeline progressively transformed a natural-language question into a compact lexical representation suitable for retrieval. Case folding and cleaning did not alter the semantic content of the example but removed representational variation that would otherwise provide no additional ranking information. Tokenization decomposed the sentence into individual lexical units, after which stopwords removal eliminated *bagaimana* while retaining *syarat*, *mendaftar*, *tugas*, and *akhir* as the principal content-bearing tokens. The final query-expansion stage introduced *pendaftaran*, *daftar*, and *registrasi*, thereby increasing the number of lexical routes through which passages discussing final-project registration could be matched. This transformation illustrates why text preprocessing and retrieval should be treated as interdependent components rather than independent procedures, since BM25 relevance scores are ultimately calculated from the lexical evidence surviving the preprocessing stage (Darmawan et al., 2025; Siregar et al., 2025).

Query expansion was deliberately applied only to user queries and not to the textual content stored in the document index. Expanding indexed passages could artificially increase the frequency of related terms and modify the corpus-level statistics on which BM25 depends, whereas query-side expansion broadens lexical matching without changing the statistical representation of the authoritative documents. This distinction is important because synonym expansion is intended to address vocabulary mismatch rather than rewrite or semantically augment the institutional corpus itself. Research on Indonesian information retrieval similarly indicates that synonym-based expansion can improve correspondence between user terminology and document terminology when the original wording differs (Suparto et al., 2025), while broader query-expansion studies demonstrate the usefulness of enriching sparse queries with additional lexical evidence before retrieval (Wang et al., 2023). The implemented design consequently preserved the lexical fidelity of the official guidebooks while allowing temporary query representations to accommodate alternative expressions commonly used by students.

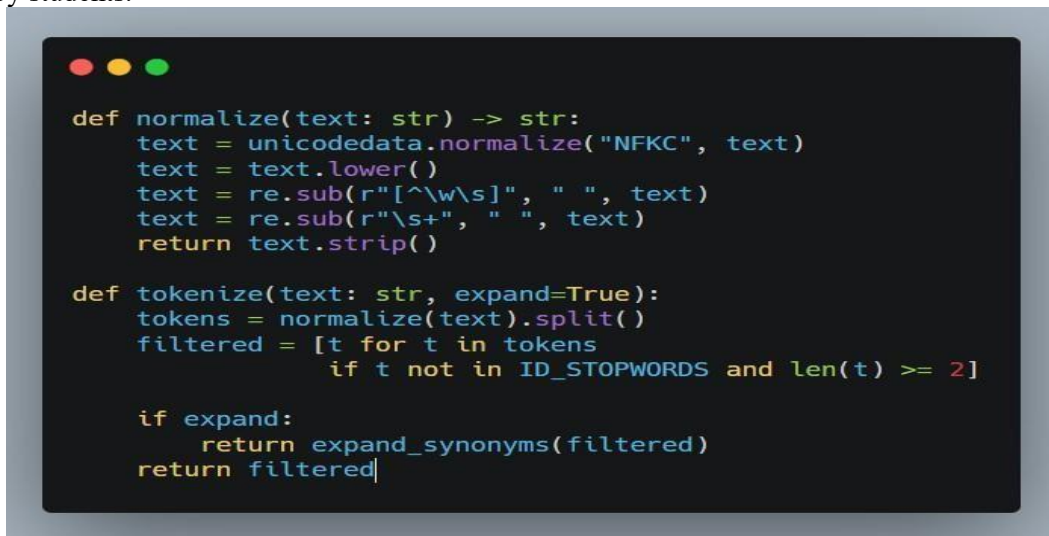


Figure 5. Chunk Formation, Paragraph Overlap, and Metadata Storage

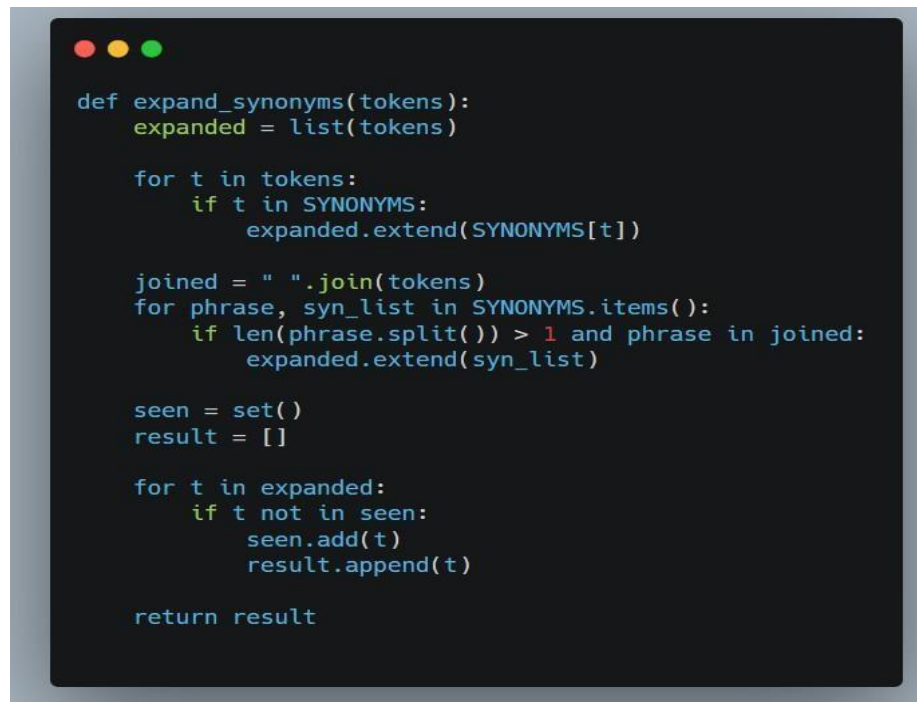
Figure 5 further illustrates that chunk formation, contextual overlap, and metadata preservation operated as an integrated preparation mechanism rather than as unrelated preprocessing operations. Each chunk retained its text, structural heading path, and identifier, while overlap preserved continuity when procedural explanations crossed a segmentation boundary. This design supports source traceability because a passage retrieved by the system can be connected to a specific structural location rather than being presented as an isolated sentence detached from its institutional origin. Source-grounded retrieval is especially relevant for question-answering environments in which procedural information should remain auditable and should not be replaced by unsupported generated statements (Karpukhin et al., 2020). The preservation of document provenance also differentiates the present architecture from conversational systems whose responses cannot easily be traced to a specific authoritative passage.

The preparation stage additionally established a distinction between body-text normalization and structural information represented through document headings. Heading information was retained because titles such as *Academic Requirements*, *Final Project Registration*, *Proposal Seminar*, and *Final Project Examination* provide topical evidence capable of differentiating passages whose body text contains overlapping procedural terminology. The subsequent indexing process could therefore exploit document structure rather than treating all tokens as contextually equivalent, while the normalized body text remained available for conventional lexical matching. This arrangement is compatible with the broader principle that retrieval performance arises from the interaction among document representation, query representation, and ranking mechanisms rather than from a ranking formula alone (Darmawan et al., 2025). It also addresses a practical characteristic of the UNISSULA corpus in which the same administrative vocabulary may recur across several stages of final-project completion but acquire different meanings according to its heading context.

The combined results of PDF reconstruction, heading-aware chunking, overlap management, metadata storage, lexical normalization, and query-specific synonym expansion show that substantial retrieval decisions were made before BM25 scoring was executed. The preparation pipeline reduced layout-related noise while maintaining procedural structure, constrained chunks to manageable retrieval units, and preserved the lexical statistics of the authoritative corpus. These findings support the view that domain-specific question-answering quality depends not only on the retrieval model but also on upstream information engineering that determines what evidence is available to that model (Bunda et al., 2025; Agustiani et al., 2026). The process also maintained a relatively lightweight and inspectable architecture, allowing individual chunks, headings, tokens, and query expansions to be examined when an unexpected result occurs rather than obscuring the retrieval process within an opaque representation. The prepared corpus consequently provides the technical and informational basis for the next analytical stage, which examines how expanded queries are scored through BM25, converted into ranked Top-5 candidates, and reconstructed into source-grounded answers.

Query Processing, BM25 Retrieval, Ranking, and Answer Presentation

Query processing constituted the operational bridge between natural-language questions submitted by students and the indexed academic-document corpus. Each incoming query passed through the same normalization sequence used during document preprocessing before being enriched through a predefined synonym-expansion mechanism. The synonym dictionary incorporated abbreviations, semantically related expressions, Indonesian morphological variants, contextual terminology, and selected multi-word phrases commonly encountered in the Informatics Engineering academic environment at UNISSULA. This mechanism was necessary because students may formulate equivalent information needs using expressions that differ substantially from the formal wording of the Final Project Guidebook and Supervision Logbook. Indonesian information-retrieval research has shown that controlled synonym expansion can broaden lexical correspondence while preserving the underlying search intention (Suparto et al., 2025), while NLP-based retrieval more generally depends on the capacity to accommodate linguistic variation between queries and stored information (Raj & Mishra, 2025).



```
def expand_synonyms(tokens):
    expanded = list(tokens)

    for t in tokens:
        if t in SYNONYMS:
            expanded.extend(SYNONYMS[t])

    joined = " ".join(tokens)
    for phrase, syn_list in SYNONYMS.items():
        if len(phrase.split()) > 1 and phrase in joined:
            expanded.extend(syn_list)

    seen = set()
    result = []

    for t in expanded:
        if t not in seen:
            seen.add(t)
            result.append(t)

    return result
```

Figure 6. Query Expansion Function

The implemented expansion function produced observable lexical changes in representative student queries without replacing the terms initially supplied by the user. The query “*Syarat TA*” was expanded into *syarat*, *persyaratan*, *ta*, *tugas*, and *akhir*, while “*Daftar Sidang*” produced *daftar*, *pendaftaran*, *mendaftarkan*, and *sidang*. The expression “*Seminar Proposal*” was represented by *seminar proposal*, *tahap*, *I*, and *I*, whereas “*Durasi Sidang*” was expanded into *durasi*, *lama*, *waktu*, *menit*, *jam*, and *sidang*. These examples demonstrate that expansion was designed to capture abbreviations, morphological variation, stage terminology, and semantically related temporal expressions rather than merely duplicate existing query terms. Query enrichment can be particularly useful for short questions because sparse formulations provide limited lexical evidence to a ranking model, a problem also addressed by contemporary query-expansion approaches (Wang et al., 2023).

Vocabulary expansion also addresses an inherent limitation of purely lexical retrieval, in which semantically related expressions may fail to interact when they share few or no surface terms. Comparative research involving BM25 and Indonesian semantic-retrieval approaches indicates that lexical and representation-based methods respond differently to vocabulary mismatch, making query formulation an important determinant of retrieval outcomes (Lasminingrum et al., 2026). The present system reduced this problem through deterministic query expansion instead of introducing learned semantic representations, enabling every additional token to remain inspectable. Such transparency is valuable for an institutional information system because unexpected results can be traced back to the original query and the terms introduced during expansion. The approach also avoids unrestricted reformulation of student questions, which helps preserve the relationship between the submitted information need and the terms subsequently used by the retrieval model.

The query-expansion mechanism was designed to deduplicate added tokens while retaining their operational order so that unnecessary repetitions did not artificially strengthen individual terms during BM25 matching. This decision is important because lexical retrieval depends on statistical evidence derived from the interaction between query terms and document terms rather than on semantic interpretation independent of token occurrence. Expansion was applied only temporarily to user queries, leaving the indexed guidebook corpus unchanged and preserving the original document-frequency statistics used by BM25. Research on preprocessing-supported BM25 applications similarly demonstrates that the effectiveness of lexical ranking can be influenced by how textual representations are produced before scoring takes place (Siregar et al., 2025). The design consequently treated expansion as a controlled mechanism for increasing possible lexical correspondence rather than as a means of modifying authoritative institutional content.

The four expansion examples also indicate that the system was adapted to the language patterns of its specific academic domain rather than using a generic vocabulary detached from final-project administration. Terms such as *TA*, *sidang*, *seminar proposal*, *pendaftaran*, and expressions related to duration are directly associated with the procedural information that students are expected to retrieve from the source documents. Domain adaptation of the query layer is important because relevance arises from the relationship between an information need and the particular context in which retrieved information will be used (Buckland & de Fremery, 2024). Academic virtual assistants likewise become practically valuable when conversational queries can be mapped to information contained in an appropriate institutional knowledge source (Amin et al., 2024). Table 2 presents the original expansion outputs implemented in the developed system.

Table 2. Example of Query Expansion Results

User Query	Expansion Result
Syarat TA	syarat, persyaratan, ta, tugas, akhir
Daftar Sidang	daftar, pendaftaran, mendaftarkan, sidang
Seminar Proposal	seminar proposal, tahap, I, 1
Durasi Sidang	durasi, lama, waktu, menit, jam, sidang

BM25Okapi was subsequently used to calculate relevance scores between each processed query and all indexed document chunks, with the implemented configuration retaining $k1 = 1.0$ and $b = 0.3$. The $k1$ parameter regulated the contribution of term frequency, whereas b controlled the extent of document-length normalization; the relatively low b value was compatible with the corpus because the preceding chunking procedure had already produced retrieval units with comparatively controlled lengths. Heading tokens received additional structural weight so that topical information encoded in the document hierarchy could contribute alongside normalized body content. BM25 combines term-frequency evidence, inverse document frequency, and document-length normalization to differentiate stronger lexical matches from weaker candidates (Darmawan et al., 2025). The manuscript's original score-calculation example reports *Academic Requirements* at 8.214, *Final Project Registration* at 7.865, *Supervisor* at 5.491, *Proposal Seminar* at 4.783, and *Examination* at 3.996, illustrating how procedurally related passages received different numerical relevance values.

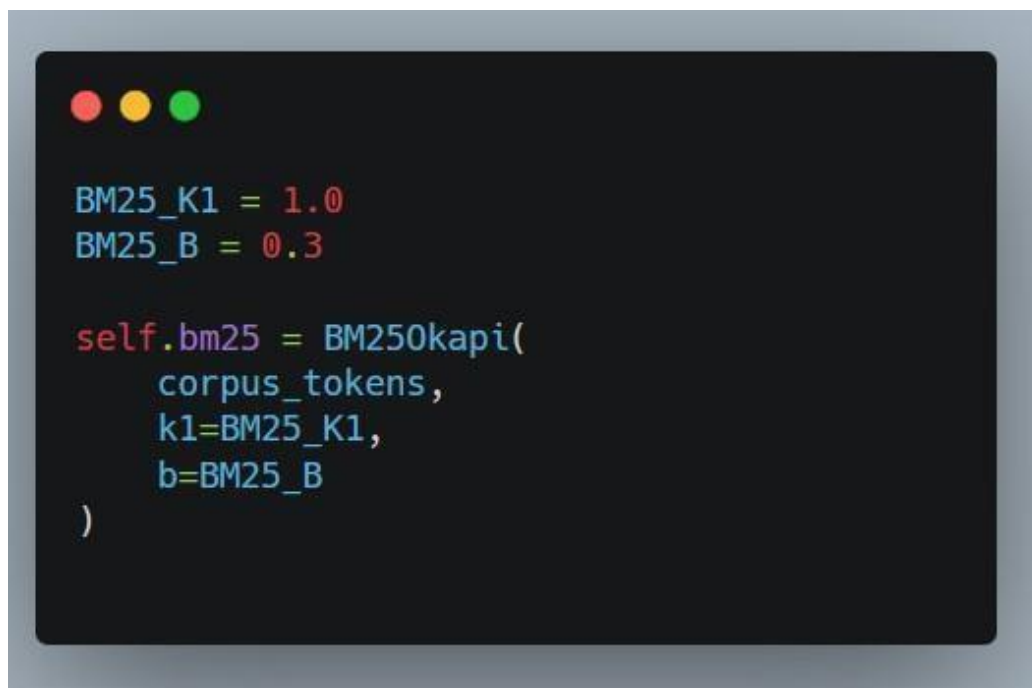


Figure 7. BM25 Configuration



```
scores = self.bm25.get_scores(q_tokens)
ranked = sorted(
    enumerate(scores),
    key=lambda x: x[1],
    reverse=True
)
```

Figure 8. BM25 Score Computation

The score distribution illustrates both the utility and the interpretive limitation of BM25 in an academic procedural corpus. A value such as 8.214 should not be interpreted as a probability that the corresponding passage is factually correct, but as relative lexical evidence for ordering that passage against competing chunks for a particular query. BM25-based search applications similarly employ graded relevance scores to prioritize textual candidates rather than treating every lexical match as equally relevant (Siregar et al., 2025). Strong surface overlap may nevertheless favor one passage over another even when the lower-scoring passage is closer to the manually defined information need, particularly when vocabulary such as *registration*, *seminar*, *supervisor*, and *examination* occurs throughout several related sections. This characteristic explains why retrieval quality in the present research was evaluated not merely through the existence of a positive BM25 score but through the ranked position ultimately assigned to the relevant source passage.

Document ranking transformed the calculated scores into an ordered Top-5 result set by eliminating duplicate query tokens, discarding chunks with scores less than or equal to zero, and sorting the remaining candidates in descending order. The original document-ranking example reports *1.4 Academic Requirements* at rank 1 with 8.214, *1.5 Final Project Registration* at rank 2 with 7.865, *2.2 Supervisor* at rank 3 with 6.547, *3.1 Proposal Seminar* at rank 4 with 5.982, and *4.2 Final Project Examination* at rank 5 with 5.401. These ranking values are retained exactly as reported in the original manuscript and constitute a separate ranking example from the earlier BM25 score-calculation illustration, which reports 5.491, 4.783, and 3.996 for the last three candidates. The source manuscript does not provide an explicit explanation for the difference between those two illustrative sets, so the present analysis preserves both sets without treating them as interchangeable measurements. Ranking remains the more practically meaningful output because users interact with an ordered result list, and Information Retrieval theory emphasizes that the position of relevant evidence influences its operational usefulness even when that evidence exists somewhere in the candidate set (Buckland & de Fremery, 2024; Moffat, 2022).

The answer-construction stage used the highest-ranked chunk as the principal source while allowing sequential chunks associated with the same heading to be combined when relevant information had been divided during the previous segmentation stage. Paragraph overlap introduced during document preparation could create repeated text at chunk boundaries, so duplicated material was removed when adjacent passages were reconstructed for presentation. The source heading was preserved in the final response, enabling users and system evaluators to identify the institutional section from which the retrieved information originated. This procedure distinguishes the architecture from systems that rely solely on generated responses because the displayed content remains anchored to textual evidence contained in official documents. Retrieval-oriented question answering similarly emphasizes the value of accessing relevant source passages before or instead of unconstrained answer generation, particularly when factual provenance is important (Karpukhin et al., 2020).

Answer delivery was implemented through a Streamlit interface that translated the retrieval pipeline into an interactive academic-information service. The interface displayed the highest-ranked passage as the primary result while retaining supporting results, their relative ordering, and their source headings for further inspection. Figure 9 presents the original search-results interface, whereas Figure 10 shows an example of a response generated after the submitted question had passed through preprocessing, query expansion, BM25 scoring, document ranking, and answer reconstruction. Conversational academic-information systems can reduce the effort required to locate administrative content when the interaction layer is connected to an appropriate institutional knowledge base (Amin et al., 2024; Mursidah et al., 2022). Comparable chatbot implementations also demonstrate that conversational access becomes useful when the underlying retrieval mechanism can connect short user expressions to an identifiable documentary source (Lawami et al., 2024).

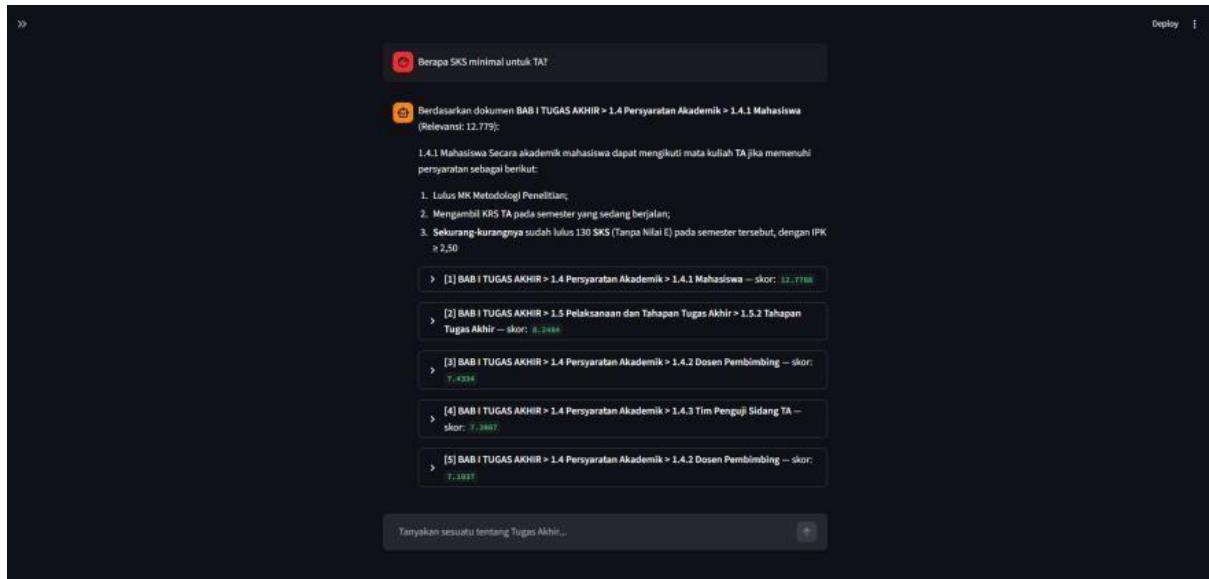


Figure 9. Search Results Interface

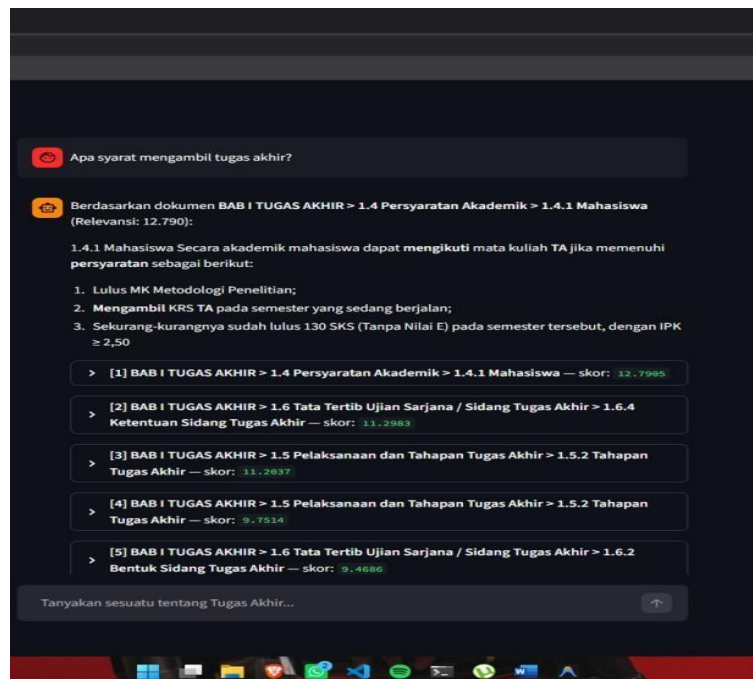


Figure 10. Example of a User Query Result

The complete retrieval workflow demonstrates that query expansion, BM25 scoring, ranking, answer reconstruction, and interface presentation operated as mutually dependent components rather than isolated computational procedures. Query expansion increased lexical correspondence, BM25 quantified the relative evidence associated with indexed passages, Top-5 ranking prioritized the strongest candidates, and the Streamlit layer exposed those results while retaining their documentary provenance. The architecture consequently remained lightweight and interpretable because retrieval errors could be investigated through expanded terms, heading paths, candidate scores, and ranking positions rather than being hidden within unrestricted text generation. This traceability is particularly important for final-project administration because incorrect information concerning registration, supervision, proposal seminars, examinations, or related academic procedures can directly affect student decisions, while domain-oriented question-answering systems require reliable access to the underlying knowledge source (Bunda et al., 2025; Karpukhin et al., 2020). The resulting retrieval output establishes the basis for the subsequent performance evaluation, where the ability of the system to locate manually mapped ground-truth passages is examined quantitatively through Recall@1, Recall@3, Recall@5, and Mean Reciprocal Rank.

Retrieval Performance, Ranking Effectiveness, and System Evaluation

System effectiveness was evaluated using 50 representative queries constructed from information contained in the official Final Project Guidebook and Supervision Logbook of the Informatics Engineering Program at UNISSULA. Each evaluation query was paired with a manually identified ground-truth passage ID representing the document segment expected to satisfy the corresponding information need. Every query was processed through the complete retrieval pipeline, including lexical normalization, synonym expansion, BM25 scoring, and Top-5 ranking, after which the position of the first relevant ground-truth passage was recorded. This evaluation design focused on retrieval effectiveness rather than linguistic fluency because the developed system was intended to locate authoritative passages instead of independently generating procedural information. Explicit relevance judgments are essential for reproducible information-retrieval evaluation because textual plausibility alone cannot establish whether a retrieved item corresponds to a predefined information need (Gienapp et al., 2024), while retrieval-oriented applications in other domains similarly require identifiable target artifacts against which ranking performance can be assessed (Alsaedi et al., 2024).

The use of manually mapped ground truth provided a controlled reference for distinguishing successful retrieval from merely lexically plausible matches. Retrieval evaluation is particularly important in the present corpus because several final-project procedures employ overlapping terminology even though they refer to different administrative stages, meaning that a passage can contain many query terms without being the intended source. Contemporary Information Retrieval research emphasizes that relevance should be interpreted in relation to the user's information need and the context represented by the target document rather than simply through surface-term overlap (Buckland & de Fremery, 2024). Evaluation collections likewise provide an essential foundation for comparing retrieval behavior systematically across queries instead of drawing conclusions from isolated examples (Lawrie et al., 2025). The selected metrics therefore captured both whether the target passage was recovered and how prominently it appeared in the ordered result list, which is consistent with the broader role of quantitative evaluation in data-driven NLP and retrieval research (Eboseremen et al., 2021; Wang, 2024).

The observed rank distribution shows that 41 of the 50 queries retrieved their designated ground-truth passage at rank 1, while eight queries first located the relevant passage at rank 2 and one query first located it at rank 3. No relevant passage was first encountered at rank 4 or rank 5, and none of the 50 evaluation queries failed to recover its ground truth within the Top-5 results. This distribution indicates that retrieval errors were not primarily caused by an inability to locate authoritative information in the corpus, but by the relative ordering of closely competing passages. Ranking position matters because the practical usefulness of a retrieved document generally decreases when users must inspect more candidates before encountering the intended evidence (Moffat, 2022). Comparative information-retrieval research also demonstrates that different retrieval methods can vary substantially in how they prioritize relevant items, making rank position a necessary component of system assessment rather than a secondary presentation concern (Ikhsan & 'Uyun, 2026).

The 41 rank-1 successes correspond to a Recall@1 of 82.00%, indicating that the system immediately presented the manually identified target passage for slightly more than four out of every five evaluated questions. The remaining 18% of queries did not represent complete retrieval failures because their relevant evidence was still available at either rank 2 or rank 3. This distinction is analytically important because top-rank failure and candidate-generation failure represent different engineering problems: the former requires improved discrimination among retrieved passages, whereas the latter would require broader improvements in corpus representation or retrieval coverage. Benchmark-oriented IR evaluation similarly treats high first-position performance as a more demanding objective than simply finding relevant evidence somewhere in a larger candidate set (Lawrie et al., 2025). The observed Recall@1 therefore establishes a strong but non-perfect first-position baseline against which later modifications to synonym expansion, heading weighting, hybrid retrieval, or reranking can be measured.

Recall@3 reached 100.00%, meaning that every manually defined relevant passage occurred within the first three retrieved candidates, while Recall@5 also reached 100.00%. The equality of these two metrics indicates that extending inspection from three to five candidates did not recover any additional ground-truth passage because all targets had already appeared within the Top-3 set. This pattern suggests that BM25 functioned effectively as a candidate generator for the evaluated institutional corpus, even when it did not consistently assign the relevant passage to the first position. High-recall candidate generation is especially valuable in multi-stage retrieval architectures because reranking mechanisms can only improve ordering when the appropriate evidence has already entered the candidate pool (Sager et al., 2025). The result therefore narrows the principal optimization target from increasing retrieval depth to improving discrimination among a small group of highly related passages.

Table 3. Retrieval Evaluation Summary

Indicator	Result
Number of evaluation queries	50
Ground truth first found at rank 1	41 queries
Ground truth first found at rank 2	8 queries
Ground truth first found at rank 3	1 query
Ground truth first found at ranks 4–5	0 queries
Recall@1	82.00%
Recall@3	100.00%
Recall@5	100.00%
Mean Reciprocal Rank (MRR)	0.9067

The Mean Reciprocal Rank of 0.9067 provides a complementary measure of how early the first correct passage appeared across the complete set of queries. The value results from 41 queries with reciprocal rank 1, eight queries with reciprocal rank 0.5, and one query with reciprocal rank approximately 0.3333, producing an aggregate value close to the theoretical maximum of 1.0. Reciprocal-rank measures are useful when retrieval tasks place particular importance on exposing the first relevant answer as early as possible rather than merely recovering it somewhere within a larger result list (Molavi & Zeinali, 2023). MRR therefore adds information that cannot be obtained from Recall@3 alone because two systems could achieve identical Top-3 coverage while differing substantially in the proportion of queries whose correct passage is ranked first. The observed value confirms that the system's perfect Top-3 recall was accompanied by a strong concentration of relevant passages in the first position rather than systematic dependence on second- or third-ranked results.

The nine cases in which the relevant passage did not occupy rank 1 reveal the principal limitation of the implemented lexical architecture. BM25 primarily exploits lexical evidence, making it possible for a competing passage with stronger term overlap or more favorable term-frequency characteristics to outrank the manually validated target even when both passages are procedurally related. The problem is especially plausible in the Final Project Guidebook because expressions associated with *registration*, *supervision*, *seminar*, *revision*, *examination*, and academic requirements may recur under several headings. Research on BM25-based retrieval confirms the importance of

term-frequency and document-level lexical representation in determining relative relevance scores (Darmawan et al., 2025; Siregar et al., 2025), while preprocessing studies demonstrate that token selection can alter the textual evidence subsequently available to retrieval algorithms (Setiawan et al., 2025; Solihin et al., 2026). The current error pattern therefore suggests that future improvement should concentrate on distinguishing semantically and structurally similar passages rather than merely introducing additional candidates.

The results also provide an empirical basis for considering more complex retrieval architectures without implying that such approaches would automatically outperform the present system. Dense passage retrieval can represent semantic correspondence beyond direct lexical overlap (Karpukhin et al., 2020), while comparative Indonesian retrieval research has reported different behaviors between BM25, baseline IndoSBERT, and fine-tuned IndoSBERT approaches (Lasminingrum et al., 2026). Hybrid and reranking pipelines offer another possibility because lexical retrieval can first generate a strong candidate set before a later component refines the ordering among competing passages (Sager et al., 2025; Schofield et al., 2025). Query reranking research similarly demonstrates that expansion and reranking can be combined to improve candidate prioritization in question-answering environments (Wei et al., 2023). Such extensions should nevertheless be evaluated against the present baseline rather than assumed to be superior, particularly because increased architectural complexity must justify itself through measurable gains in first-rank discrimination while retaining the interpretability and source traceability of the current system.

From an institutional perspective, the combination of Recall@3 and Recall@5 at 100.00% is important because no evaluated information need resulted in the complete absence of its mapped authoritative passage from the returned candidates. Academic information assistants are useful only when their conversational interface remains connected to an appropriate knowledge source rather than becoming detached from official administrative information (Amin et al., 2024; Mursidah et al., 2022). The direct-passage design employed in this study retains this connection because retrieved answers remain associated with their document headings and identifiers, while the Streamlit interface functions primarily as the presentation layer. Controlled synonym expansion further accommodates variation between student expressions and formal institutional vocabulary without rewriting the underlying corpus (Suparto et al., 2025; Wang et al., 2023). The resulting architecture therefore combines relatively strong ranking performance with auditability, a characteristic that remains relevant as retrieval research increasingly explores more autonomous or agentic information-access architectures (Zhang & Lin, n.d.).

Overall, the evaluation establishes a reproducible quantitative baseline for BM25-based question answering within a small, structured, Indonesian academic-information corpus. Recall@1 of 82.00% demonstrates strong immediate retrieval, while Recall@3 and Recall@5 of 100.00% show that every ground-truth passage remained available within a very small candidate set, and the MRR of 0.9067 confirms that relevant evidence was generally concentrated near the top of the ranking. The findings are consistent with the broader observation that retrieval quality emerges from the combined effects of corpus preparation, chunking, preprocessing, query expansion, indexing, scoring, and ranking rather than from a single algorithmic component (Agustiani et al., 2026; Fahreza & Anwar, 2026). Lexical retrieval remains attractive for this application because the corpus is controlled, the vocabulary is procedurally distinctive, and individual scores and passages can be inspected, while NLP-supported retrieval provides the broader computational framework through which natural-language information needs are mapped to documentary evidence (Raj & Mishra, 2025). The principal opportunity for future development is consequently not to replace the existing architecture without evidence, but to determine whether semantic, hybrid, reranking, or more advanced retrieval mechanisms can move the eight rank-2 cases and one rank-3 case to the first position without sacrificing the complete Top-3 coverage, transparency, maintainability, and institutional source fidelity already achieved.

CONCLUSION

This study demonstrates that a lightweight BM25-based question-answering architecture can effectively support the retrieval of authoritative final-project information for Informatics Engineering students at Universitas Islam Sultan Agung. The integration of document reconstruction, heading-aware chunking, paragraph overlap, lexical preprocessing, query-specific synonym expansion,

heading-weighted indexing, and source-grounded answer presentation produced a retrieval process that remained transparent and traceable to official institutional documents. Evaluation using 50 manually mapped ground-truth queries resulted in Recall@1 of 82.00%, Recall@3 of 100.00%, Recall@5 of 100.00%, and a Mean Reciprocal Rank of 0.9067. Forty-one queries placed the relevant passage at rank 1, eight at rank 2, and one at rank 3, while no relevant passage first appeared at ranks 4 or 5. These results indicate that the system was highly effective in recovering relevant academic information within a compact candidate set.

The principal limitation lies in first-rank discrimination among passages containing overlapping procedural terminology rather than in the failure to retrieve relevant information. The perfect Top-3 coverage suggests that future development should focus on improving the ordering of closely related candidates without sacrificing the transparency, maintainability, and source fidelity provided by the current lexical architecture. Hybrid retrieval, semantic representations, or reranking mechanisms may be evaluated against the established BM25 baseline to determine whether they can improve the nine queries whose relevant passages did not occupy the first position. Broader testing with additional institutional documents and more diverse student queries would also strengthen the assessment of generalizability. The present findings provide a reproducible foundation for developing interpretable academic question-answering systems in small, structured, domain-specific information environments.

REFERENCES

- Agustiani, A. D., Putri, S. M., Hutahaean, J., Rizqi, M., Alifi, M. R., & Hodijah, A. (2026). Evaluating RAG performance on small language models for low-resource devices through chunking and retrieval methods. *Jurnal Online Informatika*, 11(1), 243–255. <https://doi.org/10.15575/join.v11i1.1733>
- Alsaedi, S., Gad-Elrab, A. A., Noaman, A., & Eassa, F. (2024). Two-level information-retrieval-based model for bug localization based on bug reports. *Electronics*, 13(2), 321. <https://doi.org/10.3390/electronics13020321>
- Amin, A., Bustomi, T., & Harianto, K. (2024). Implementasi chatbot sebagai asisten virtual untuk membantu mahasiswa dalam mendapatkan informasi akademik. *Sebatik*, 28(2), 1–12. <https://doi.org/10.46984/sebatik.v28i2.0000>
- Atmojo, T. T., & Kunang, Y. N. (2023). Machine learning-based e-archive for archives management of South Sumatra Province. *Journal of Information Systems and Informatics*, 5(4), 1491–1507. <https://doi.org/10.51519/journalisi.v5i4.566>
- Buckland, M. K., & de Fremery, W. (2024). Relevance, works and context engineering. *Proceedings from the Document Academy*, 11(2), Article 5. <https://doi.org/10.35492/docam/11/2/5>
- Bunda, Y. P., Siregar, M. F., Sains, F., Studi, P., Informasi, S., Tjut, U., & Dhien, N. (2025). Pengembangan Question Answering System Menggunakan Large Language Model UU TNI 2025. *Jurnal Pendidikan Dan Teknologi Indonesia*, 5(10), 3041-3050. <https://doi.org/10.52436/1.jpti.1396>
- Darmawan, I., Arifin, M. N., Ubaidi, U., Ramadhani, N., & Dewi, N. P. (2025). Algoritma Okapi Bm25 Dalam Sistem Pencarian Informasi Berbasis Teks. *Insand Comtech: Information Science and Computer Technology Journal*, 10(1), 30-37. <https://doi.org/10.53712/jic.v10i1.2649>
- Eboseremen, B., Adebayo, A., Essien, I., Afuwape, A., Soneye, O., & Ofori, S. (2021). The role of natural language processing in data-driven research analysis. *International Journal of Multidisciplinary Research and Growth Evaluation*, 2(1), 935-942. <https://doi.org/10.54660/IJMRGE.2021.2.1.935-942>
- Fahreza, R. A., & Anwar, M. (2026). Routing struktural adaptif: Arsitektur hybrid chunking untuk RAG pada dokumen pembelajaran teknik informatika. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6(Nomor Jurnal), 1356–1365. <https://doi.org/10.58578/masaliq.v6i3.10134>
- Gienapp, L., Scells, H., Deckers, N., Bevendorff, J., Wang, S., Kiesel, J., Syed, S., Fröbe, M., Zuccon, G., Stein, B., Hagen, M., & Potthast, M. (2024). Evaluating generative ad hoc information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and*

- Development in Information Retrieval* (pp. 1916–1929). Association for Computing Machinery. <https://doi.org/10.1145/3626772.3657849>
- Ikhsan, M. F., & 'Uyun, S. . . (2026). Analisis Perbandingan Metode Information Retrieval (IR) Pada Implementasi Pencarian Ayat Al-Qur'an. *Journal Software, Hardware and Information Technology*, 6(1), 83–92. <https://doi.org/10.24252/shift.v6i1.272>
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- Lasminingrum, D. P., Puspaningrum, E. Y., & Mulyo, B. M. (2026). Perbandingan metode lexical dan semantic retrieval: BM25, IndoSBERT baseline, dan IndoSBERT fine-tuned pada pencarian dokumen berbahasa Indonesia. *Prosiding Seminar Nasional Penelitian dan Pengabdian kepada Masyarakat LPPM Universitas 'Aisyiyah Yogyakarta*, 4. <https://proceeding.unisayogya.ac.id/index.php/proseminaslppm/article/view/2088>
- Lawami, R. H., Harahap, N. S., & Iskandar, I. (2024). Pembuatan Chatbot Telegram untuk Layanan Pencarian Al-Quran. *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, 5(2), 1846–1857. <https://doi.org/10.35870/jimik.v5i2.761>
- Lawrie, D., Mayfield, J., Yang, E., Yates, A., MacAvaney, S., Pradeep, R., ... & Soldani, L. (2025). Neucirbench: A modern evaluation collection for monolingual, cross-language, and multilingual information retrieval. *arXiv preprint arXiv:2511.14758*. <https://doi.org/10.48550/arXiv.2511.14758>
- Mallia, A., Mackenzie, J., Suel, T., & Tonellotto, N. (2022). Faster learned sparse retrieval with guided traversal. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery. <https://doi.org/10.1145/3477495.3531774>
- Moffat, A. (2022). Batch evaluation metrics in information retrieval: Measures, scales, and meaning. *IEEE Access*, 10, 105564–105577. <https://doi.org/10.1109/ACCESS.2022.3211668>
- Molavi, M., & Zeinali, H. (2023, July). SLT at SemEval-2023 Task 1: Enhancing visual word sense disambiguation through image text retrieval using BLIP. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 1921–1925). <https://doi.org/10.18653/v1/2023.semeval-1.264>
- Mursidah, E., Ambarwati, L., & Karima, F. A. (2022). Implementasi Chatbot layanan informasi pendaftaran mahasiswa baru program pascasarjana Departemen Teknik Informatika ITS. *J. Ilm. Nero*, 7(1), 43–52. <https://doi.org/10.21107/nero.v7i1.276>
- Raj, M., & Mishra, N. (2025). Exploring new approaches for information retrieval through natural language processing. *arXiv preprint arXiv:2505.02199*. <https://doi.org/10.48550/arXiv.2505.02199>
- Sager, P. J., Kamaraj, A., Grewe, B. F., & Stadelmann, T. (2025). Deep Retrieval at CheckThat! 2025: identifying scientific papers from implicit social media mentions via hybrid retrieval and re-ranking. *arXiv preprint arXiv:2505.23250*. <https://doi.org/10.48550/arXiv.2505.23250>
- Schofield, J., Tian, S., Truong, H. T. T., & Heil, M. (2025). DS@ GT at CheckThat! 2025: exploring retrieval and reranking pipelines for scientific claim source retrieval on social media discourse. *arXiv preprint arXiv:2507.06563*. <https://doi.org/10.48550/arXiv.2507.06563>
- Setiawan, A., Abidin, Z., & Imammudin, M. (2025). Impact of preprocessing on Indonesian extractive summarization using LexRank, TextRank, DivRank, and cosine similarity. *G-Tech: Jurnal Teknologi Terapan*, 9(4), 2311–2321. <https://doi.org/10.70609/g-tech.v9i4.8306>
- Siregar, M. E., Benedicta, L., Adrielvino, N. A., Park, S. B., Gading, J., Boulevard, S., Tower, N., Sangereng, C., & Dua, K. K. (2025). Pengembangan sistem pencarian resep makanan dengan implementasi text preprocessing dan BM25. *Jurnal Nasional Komputasi dan Teknologi Informasi*, 8(3), 1458–1465. <https://doi.org/10.32672/jnkti.v8i3.9137>
- Solihin, A. A., Utomo, F. S., & Barkah, A. S. (2026). Impact of Stopword Variation on Qur'anic Text Classification using Support Vector Machine and Backpropagation. *Journal of Innovation Information Technology and Application (JINITA)*, 8(1), 173–183. <https://doi.org/10.35970/jinita.v8i1.3069>

- Suparto, A., Clement, M. J., Rahman, A., & Irsyad, H. (2025). Implementasi Preprocessing Dan Synonym Expansion Untuk Sistem Temu Kembali Berita Bahasa Indonesia. *Jurnal Software Engineering and Computational Intelligence*, 3(01), 10-17. <https://doi.org/10.36982/jseci.v3i01.5405>
- Wang, L., Yang, N., & Wei, F. (2023, December). Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9414-9423). <https://doi.org/10.18653/v1/2023.emnlp-main.585>
- Wang, S. (2024). Exploring the fusion of natural language processing and information retrieval. *Applied and Computational Engineering*, 76, 27-32. <https://doi.org/10.54254/2755-2721/76/20240559>
- Wei, Y. C., Li, F. S., & Glass, J. (2023). Expand, rerank, and retrieve: Query reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 12131–12147). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.768>
- Zhang, W., & Lin, J. (n.d.). *Agentic Information Retrieval*. Association for Computing Machinery.