Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Knowledge Grounded Conversational Access to Heterogeneous Institutional Documents via OCR Enabled Hybrid RAG

Veerababu Reddy

rveerababu_vlits@vignan.ac.in

Vignan's Lara Institute of Technology and Science

Krishnaveni Kaki

Vignan's Lara Institute of Technology and Science

Bharath Kesineni

Vignan's Lara Institute of Technology and Science

Iswarya Annapureddy

Vignan's Lara Institute of Technology and Science

Harini Kanna

Vignan's Lara Institute of Technology and Science

Gargesh Chalamala

Vignan's Lara Institute of Technology and Science

Research Article

Keywords: Conversational information systems, Retrieval augmented generation, Optical character recognition, Context aware information retrieval, Knowledge grounded conversational agents, Large language models

Posted Date: April 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9287246/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Knowledge Grounded Conversational Access to Heterogeneous Institutional Documents via OCR Enabled Hybrid RAG

Veerababu Reddy[1*], Krishnaveni Kaki[1], Bharath Kesineni[1], Iswarya Annapureddy[1], Harini Kanna[1], Gargesh Chalamala[1]

[1]Vignan's Lara Institute of Technology and Science, Vadlamudi, Andhra Pradesh, 522213, India.

*Corresponding author(s). E-mail(s): rveerababu_vlits@vignan.ac.in;
Contributing authors: krishnaveni3075@gmail.com;
kesinenibharath159@gmail.com; annapureddyiswarya@gmail.com;
harinikanna406@gmail.com; gargesh2244@gmail.com;

Abstract

Academic institutions generate large volumes of information such as academic regulations, examination schedules, policies, circulars, and administrative notices. Conversational systems are increasingly used to help students and staff access this information quickly through natural language queries. However, many existing conversational systems rely mainly on large language models that generate responses based on internal model knowledge. Because these systems are not fully grounded in institutional documents, they may produce inaccurate or hallucinated responses and often fail to effectively use heterogeneous document sources such as scanned notices, PDFs, and legacy records. To address these limitations, this paper proposes a knowledge centric conversational framework for reliable academic and campus guidance. The proposed approach integrates optical character recognition (OCR), hybrid knowledge retrieval, and retrieval augmented generation (RAG) to ensure that responses are grounded in authoritative institutional knowledge. The framework first converts scanned and image based documents into machine readable text using OCR and preprocessing techniques. The extracted text is segmented into semantic knowledge fragments and represented using vector embeddings. A hybrid retrieval mechanism combining semantic similarity search and keyword matching retrieves relevant knowledge fragments, which are then used by a retrieval augmented language model. A rule

based and reasoning module further validates the generated responses. The system is evaluated on a heterogeneous institutional document dataset containing structured, unstructured, and scanned documents. Experimental results show that the proposed framework achieves 89.6% response accuracy, 0.86 precision, 0.84 recall, and an F1 score of 0.85, while reducing the hallucination rate to 4.1%, demonstrating improved reliability and contextual accuracy compared with baseline conversational systems. The implementation in this study are publicly available at DOI: https://doi.org/10.5281/zenodo.19230595.

Keywords: Conversational information systems; Retrieval augmented generation; Optical character recognition; Context aware information retrieval; Knowledge grounded conversational agents; Large language models.

1 Introduction

Academic institutions operate as knowledge intensive environments in which regulations, academic policies, procedural guidelines, schedules, and administrative notifications are continuously generated and updated. Access to this institutional knowledge is essential for students, faculty members, and administrative staff to perform decision making tasks and navigate institutional processes effectively [1]. However, in practice, institutional information is distributed across heterogeneous and partially unstructured sources, including digital documents, semi structured records, scanned notices, and legacy image based files [8]. Such fragmentation creates a significant knowledge access challenge that users must manually locate, interpret, and reconcile information from multiple sources under time constraints, often without guarantees of correctness, completeness, or contextual relevance [7, 9]. From a data and knowledge engineering perspective, this challenge extends beyond traditional document storage or search. It requires mechanisms that support end to end knowledge acquisition, representation, retrieval, and reasoning, enabling reliable and context aware information access.

Conversational interfaces have recently emerged as a promising solution for interacting with [2, 5]. Advances in large language models (LLMs) have enabled conversational agents capable of generating fluent and context sensitive responses [10]. Nevertheless, these models remain largely parametric and model centric, meaning that responses are generated primarily from internal representations rather than explicitly grounded in authoritative institutional knowledge sources [3]. As a result, generated responses may be outdated, incomplete, or unsupported by verifiable documents [4]. This limitation is particularly problematic in academic guidance scenarios where correctness, traceability, and policy compliance are critical.

Retrieval Augmented Generation has emerged as an effective paradigm for grounding language model outputs in external knowledge sources [6]. By retrieving relevant content and conditioning response generation on this retrieved context, RAG reduces reliance on parametric memory and improves factual consistency [20]. However, most existing RAG based conversational systems assume the availability of clean and machine readable corpora and treat knowledge acquisition, indexing, retrieval, and

response generation as loosely coupled stages [12]. In academic environments, a significant portion of institutional knowledge exists in scanned or image based documents, requiring robust Optical Character Recognition and preprocessing before semantic retrieval becomes feasible [11, 21].

Furthermore, many conversational systems lack explicit mechanisms for validating generated responses against institutional constraints, leading to residual hallucinations and limited reliability [13, 19]. These limitations reveal a methodological gap in current conversational information systems [14, 15]. Existing approaches do not provide a unified knowledge lifecycle that tightly integrates heterogeneous knowledge acquisition, semantic representation, context aware retrieval, grounded generation, and response validation [14]. As a consequence, institutional knowledge remains fragmented, document coverage is incomplete, and the traceability of generated responses is limited [16].

To address this gap, this work proposes a knowledge centric conversational information system for real time academic and campus guidance [17]. The proposed framework models conversational academic guidance as an end to end grounded knowledge access process [18]. It integrates OCR enabled knowledge acquisition, metadata enriched semantic representation, hybrid retrieval combining semantic similarity and keyword based search, and retrieval augmented response generation [19]. In addition, a reasoning and validation module ensures that generated responses remain consistent with retrieved institutional knowledge [21]. By preserving the flow of knowledge from source documents to generated responses, the framework enables context aware, traceable, and reliable conversational interaction over heterogeneous institutional knowledge sources [22].

1.1 Research Questions and Hypotheses

To evaluate the effectiveness of the proposed knowledge centric conversational information system, this study investigates the following research questions:

- **RQ1:** How can heterogeneous institutional documents, including scanned and unstructured sources, be effectively acquired and represented to support semantic retrieval in conversational information systems?
- **RQ2:** Does integrating OCR based knowledge acquisition with retrieval augmented generation improve the accuracy and contextual relevance of conversational responses compared with model centric conversational systems?
- **RQ3:** To what extent does the incorporation of reasoning and validation mechanisms reduce hallucinations and improve the reliability and traceability of responses in real time academic guidance scenarios?

Based on these research questions, the following hypotheses are formulated:

- **H1:** OCR enabled semantic representation and vector based indexing of heterogeneous institutional documents improve the retrieval of contextually relevant knowledge.
- **H2:** Retrieval augmented generation grounded in institutional knowledge produces more accurate and context consistent responses than standalone LLM based conversational models.

- **H3:** Integrating reasoning and validation constraints reduces hallucinated responses and improves the trustworthiness of conversational academic guidance.

1.2 Contributions

This work contributes to the field of data and knowledge engineering through both methodological and empirical advancements.

1.2.1 Methodological Contributions

- A unified knowledge centric conversational framework integrating acquisition, representation, retrieval, generation, and validation.
- OCR enabled knowledge acquisition for incorporating scanned and image based institutional documents into semantic representations.
- A grounded reasoning and validation mechanism ensuring alignment between generated responses and authoritative knowledge sources.

1.2.2 Empirical Contributions

- Lifecycle oriented evaluation analyzing accuracy, hallucination reduction, contextual relevance, and system latency.
- Ablation based analysis demonstrating that performance improvements arise from integrated system design rather than isolated components.

The remainder of this paper is organized as follows. Section 2 reviews related work on conversational systems, knowledge retrieval, and document processing, and highlights the research gap. Section 3 presents the proposed knowledge centric framework along with its formulation. Section 4 describes the methodology and implementation details. Section 5 outlines the experimental setup, followed by results and discussion in Section 6. Finally, Section 7 concludes the paper and presents future research directions.

2 Related Work

The development of conversational information systems for academic guidance intersects multiple research areas, including knowledge centric information systems, conversational question answering, document understanding, retrieval augmented generation, and trustworthy artificial intelligence. This section reviews existing research across these domains from a data and knowledge engineering perspective and identifies the methodological gap addressed in this study.

2.1 Knowledge Centric Information Systems in Academic Environments

Knowledge centric information systems model organizational information as structured, semantically interpretable, and reusable knowledge assets rather than as isolated documents [27]. This paradigm has been widely adopted in data and knowledge engineering to support complex information access in knowledge intensive

domains [4]. Academic institutions represent a particularly relevant example of such environments, as they continuously generate large volumes of regulations, procedural guidelines, academic schedules, and administrative notifications. Effective use of this information requires not only storage and retrieval but also contextual interpretation and traceability. Prior studies have shown that semantic representation and explicit knowledge modeling significantly improve retrieval effectiveness and enable context aware query processing [21]. By transforming documents into structured knowledge units, these approaches support reasoning over institutional content and facilitate knowledge reuse across multiple applications.

However, most existing implementations assume homogeneous and machine readable data sources. In practice, institutional knowledge is distributed across heterogeneous formats, including scanned documents, semi structured PDFs, and legacy image based records. Moreover, knowledge acquisition, indexing, and retrieval processes are often treated as loosely coupled stages. This fragmentation limits the ability to maintain an end to end knowledge flow from source documents to generated responses, thereby reducing the effectiveness of real time conversational information access [29].

2.2 Conversational Information Systems for Academic Guidance

Conversational interfaces have emerged as an intuitive mechanism for accessing institutional information. Early conversational systems relied on rule based dialogue management and predefined question answer repositories. Although such systems provided basic information access, their scalability and ability to handle complex or context dependent queries were limited. More recent approaches employ neural language models to enable open ended interaction and multi turn dialogue, significantly improving linguistic fluency and user engagement. Despite these advances, many conversational systems remain model centric [2].

Knowledge is implicitly encoded within model parameters rather than explicitly retrieved from authoritative sources. Consequently, the origin of generated responses cannot be traced to institutional documents, making procedural verification difficult. In academic guidance scenarios, where policies, deadlines, and regulations frequently change, this lack of grounding can lead to factual inconsistencies and reduced user trust. Furthermore, conversational interfaces are rarely integrated with structured knowledge acquisition pipelines, which limits systematic reuse of institutional knowledge and restricts the ability of these systems to provide policy compliant and context aware guidance [1].

2.3 OCR Based Knowledge Acquisition from Heterogeneous Documents

A significant portion of institutional knowledge exists only in scanned notices, printed circulars, and legacy records. Optical Character Recognition technologies enable the conversion of such documents into machine readable text and have been extensively

studied in document analysis and digital archiving [25]. Although modern OCR systems achieve high recognition accuracy under controlled conditions, their integration into knowledge centric conversational systems remains limited.

In many architectures, OCR is treated merely as a preprocessing step, and the extracted text is not systematically incorporated into semantic knowledge representations. Additionally, the impact of OCR extraction noise on retrieval performance is rarely considered. As a result, a substantial portion of institutional knowledge remains inaccessible to conversational systems [21]. This limitation is particularly significant in academic environments, where important procedural information is often available only in scanned or image based documents.

2.4 Retrieval Augmented Generation for Knowledge Grounded Conversational Systems

Retrieval Augmented Generation has been proposed as a mechanism for grounding language model outputs in external knowledge sources [6]. By retrieving relevant passages and conditioning response generation on this contextual information, RAG reduces reliance on parametric memory and improves factual consistency [22, 30]. The RAG paradigm has been successfully applied in question answering, enterprise search, and knowledge management systems.

However, most existing RAG architectures assume the availability of clean, static, and homogeneous corpora. They typically do not address the acquisition of knowledge from heterogeneous and dynamically evolving sources. Furthermore, the retrieval and generation stages are often loosely coupled, and the selection of retrieved context is not constrained by domain specific requirements [23]. In knowledge intensive environments such as academic institutions, these limitations can lead to reduced semantic recall, weaker contextual alignment, and limited traceability between generated responses and authoritative documents [24].

2.5 Reasoning and Validation in Trustworthy Conversational Systems

The reliability of conversational systems depends not only on effective retrieval but also on the ability to verify generated responses against domain constraints. Recent research in trustworthy artificial intelligence has explored hybrid neural symbolic approaches that combine language models with logical reasoning and rule based validation. These approaches have demonstrated potential for reducing hallucinations and improving interpretability [26].

However, in many existing conversational architectures, reasoning mechanisms are applied as post generation filtering processes rather than as integral components of the retrieval generation pipeline. As a result, generated responses are not systematically constrained by institutional policies or procedural rules, and validation processes often lack direct linkage to the retrieved knowledge [28]. Consequently, the end to end knowledge flow remains weakly controlled, limiting the ability of conversational systems to provide traceable and policy compliant guidance.

2.6 Research Gap

Current conversational information systems lack a unified knowledge lifecycle integrating heterogeneous acquisition, semantic representation, context aware retrieval, grounded generation, and validation [21]. Existing approaches rely on clean corpora, model centric representations, and loosely coupled pipelines, resulting in limited coverage, low semantic recall, and poor traceability. These limitations are critical in academic guidance, where correctness, relevance, and verifiability are essential. As shown in Table 1, significant gaps persist across grounding, coverage, retrieval, and validation. Hence, a knowledge centric framework is required for end to end grounded conversational access.

Table 1 Comparative Analysis of Existing Approaches and the Proposed Knowledge Centric Framework

Category	**Existing Approach**	**Limitations**	**Evaluation Parameters**	**Proposed Knowledge Centric Framework**
System Paradigm	Model centric conversational systems relying on parametric knowledge	Weak grounding in authoritative sources and limited traceability	Knowledge grounding, transparency	Knowledge centric conversational framework with explicit source alignment
Knowledge Acquisition	Clean text corpora and predefined FAQs	Inability to incorporate scanned, image based, and legacy documents	Document coverage, ingestion robustness	OCR enabled acquisition from heterogeneous institutional documents
Knowledge Representation	Plain text storage or keyword based indexing	Loss of semantic relationships and contextual meaning	Semantic fidelity, contextual coherence	Semantic embedding based representation with metadata enrichment
Retrieval Mechanism	Keyword based or single mode dense retrieval	Low recall for semantically similar and context dependent queries	Precision, recall, retrieval depth	Hybrid dense and keyword retrieval with rank fusion
LLM Integration	Direct response generation without external grounding	High hallucination rate and lack of source attribution	Faithfulness, correctness	Retrieval augmented generation grounded in authoritative knowledge
Reasoning and Validation	Minimal or post hoc validation mechanisms	No enforcement of institutional rules or procedural correctness	Consistency, policy compliance	Integrated rule based and neuro symbolic validation
System Integration	Fragmented and loosely coupled pipelines	Reduced scalability, higher latency, and limited traceability	Scalability, latency, lifecycle coherence	Unified end to end knowledge lifecycle architecture

3 Proposed Framework

This section presents the proposed knowledge-centric conversational framework designed to support reliable conversational access to heterogeneous institutional

knowledge. The framework models the problem as an end-to-end knowledge lifecycle that integrates document acquisition, semantic representation, retrieval, response generation, and validation within a unified architecture.

Unlike model-centric conversational systems that rely solely on parametric knowledge embedded in language models, the proposed framework explicitly grounds conversational responses in authoritative institutional documents. The architecture incorporates OCR-based document processing, hybrid semantic retrieval, retrieval-augmented response generation, and reasoning-driven validation to ensure accuracy, traceability, and contextual consistency.

The overall architecture of the proposed framework is illustrated in Figure. 1. The framework is organized into two primary pipelines: the *offline knowledge ingestion pipeline*, responsible for acquiring, preprocessing, and structuring institutional knowledge, and the *online conversational query pipeline*, responsible for real time user interaction, retrieval, and response generation. This separation enables scalable knowledge management while maintaining efficient and low latency conversational responses.

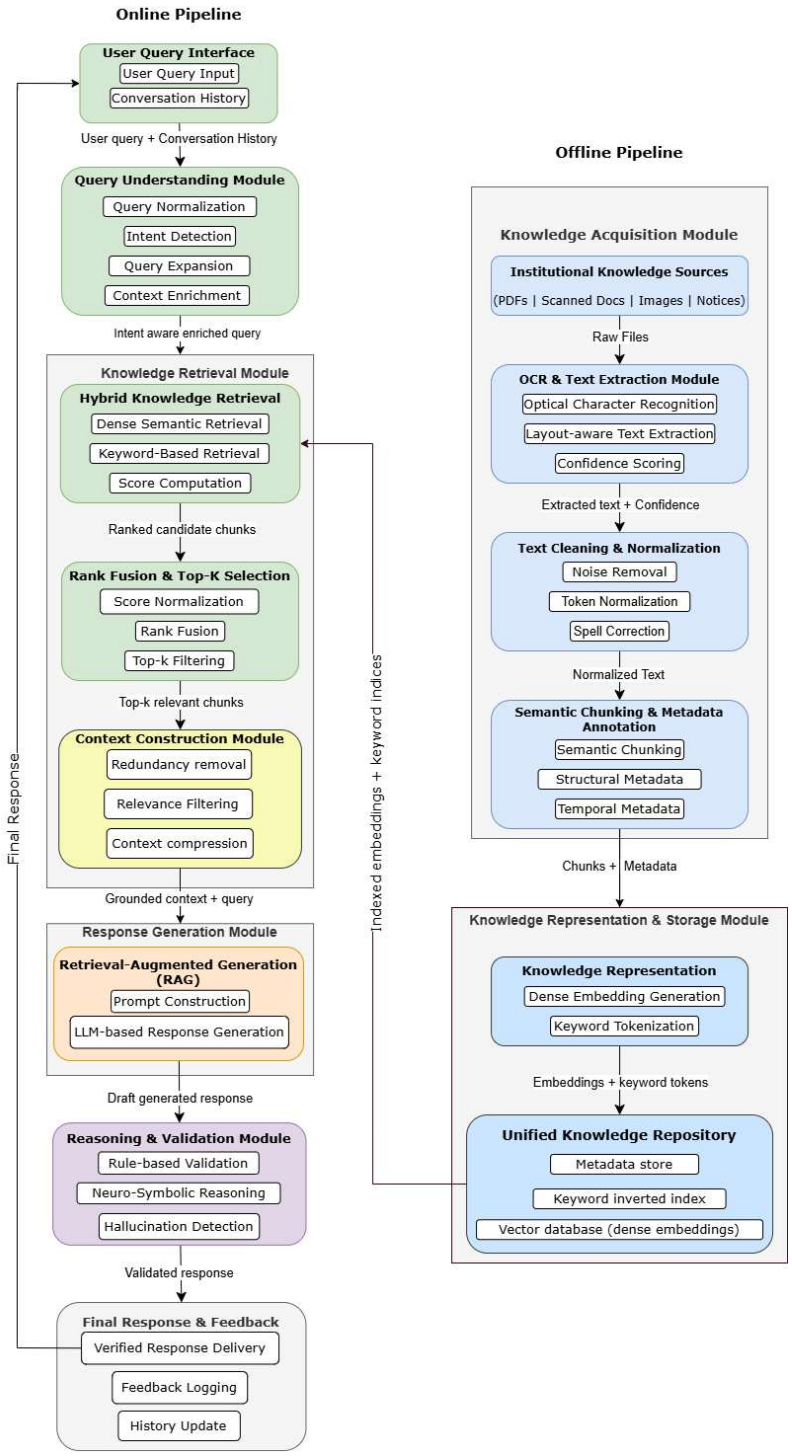

Fig. 1 Architecture of the proposed knowledge centric conversational framework for OCR enabled institutional knowledge acquisition, hybrid retrieval, and retrieval augmented response generation.

Algorithm 1 End to End Knowledge Ingestion, Retrieval, Reasoning, and Conversational Response Generation

Require: Institutional document set $D = \{d_1, d_2, \ldots, d_n\}$, user query at time t: q_t
Ensure: Validated, context aware system output O_t

Acquire institutional documents D
if documents are scanned or image based **then**
    Extract text using OCR: $T \leftarrow f_{\text{ocr}}(D)$
else
    Read textual content directly
end if
Normalize extracted text: $\hat{T} \leftarrow N(T)$
Segment text into semantic chunks: $C \leftarrow \text{Chunk}(\hat{T})$
Generate embeddings: $v_i \leftarrow E(c_i)$
Store embeddings and metadata in knowledge repository V
Receive user query q_t
Perform query preprocessing: $\hat{q}_t \leftarrow f_{\text{context}}(q_t)$
Perform dense semantic retrieval: $R_d \leftarrow \text{DenseRetrieve}(\hat{q}_t, V)$
Perform keyword retrieval: $R_k \leftarrow \text{KeywordRetrieve}(\hat{q}_t, D)$
Combine results: $R \leftarrow R_d \cup R_k$
Rank retrieved fragments: $R_t \leftarrow \text{Rank}(R)$
Construct grounded prompt: $P_t \leftarrow \langle \hat{q}_t, R_t \rangle$
Generate response: $G_t \leftarrow f_{\text{llm}}(P_t)$
Apply rule based validation: $G'_t \leftarrow f_{\text{rule}}(G_t)$
Apply neuro symbolic reasoning: $O_t \leftarrow f_{\text{neuro}}(G'_t)$
return O_t

The proposed framework implements an end to end pipeline that integrates institutional knowledge ingestion, hybrid retrieval, reasoning, and response generation. As shown in Algorithm 1, the system processes user queries through OCR enabled document handling, semantic chunking, and embedding based indexing. The enriched query is resolved using a combination of dense semantic and keyword based retrieval to construct a grounded prompt. Finally, the generated response undergoes rule based validation and neuro symbolic reasoning to ensure contextual accuracy and reliability.

3.1 System Architecture Overview

The proposed framework follows a two stage architecture consisting of an offline knowledge ingestion pipeline and an online query processing pipeline. The offline pipeline is responsible for acquiring and structuring institutional knowledge from heterogeneous sources such as digital documents, PDFs, scanned notices, and image based records. These documents are processed using OCR and preprocessing techniques to extract machine readable text. The extracted content is then segmented into semantic fragments and transformed into vector embeddings that are stored in a unified knowledge repository.

The online pipeline handles real time user queries and conversational interaction. User queries are first normalized and enriched using conversational context. The system then retrieves relevant knowledge fragments using a hybrid retrieval strategy that combines semantic similarity search and keyword based retrieval. Retrieved fragments are incorporated into a retrieval augmented generation process to produce grounded conversational responses. Finally, reasoning and validation mechanisms verify the consistency of generated responses with institutional policies. By tightly integrating knowledge acquisition, representation, retrieval, and reasoning within a unified architecture, the framework ensures that conversational responses remain traceable to authoritative institutional sources.

3.2 Knowledge Acquisition via OCR

Institutional knowledge is often distributed across heterogeneous document formats including structured digital records, PDF files, scanned notices, and image based documents. To integrate these diverse sources into the conversational system, the framework incorporates an OCR enabled knowledge acquisition module.

Let d denote a scanned or image based institutional document. The OCR extraction process converts the document into machine readable text:

$$T = f_{\text{ocr}}(d) \tag{1}$$

where d represents the input document, $f_{\text{ocr}}(\cdot)$ denotes the OCR extraction function, and T represents the extracted textual content. This process enables the system to incorporate legacy and non editable documents into the knowledge repository, ensuring comprehensive coverage of institutional information.

3.3 Text Normalization and Preprocessing

The textual content extracted through OCR may contain noise, layout artifacts, and recognition errors. Therefore, a preprocessing stage is applied to normalize and clean the extracted text before further processing. The normalization process is represented as:

$$\hat{T} = N(T) \tag{2}$$

where T is the raw extracted text, $N(\cdot)$ represents the normalization function, and $\hat{T}$ denotes the normalized textual content. The normalization process includes operations such as punctuation removal, token standardization, spelling correction, and formatting normalization. These steps improve the consistency of textual data and facilitate downstream semantic processing.

3.4 Semantic Chunking and Metadata Annotation

To support fine grained retrieval, the normalized text is segmented into semantically coherent knowledge fragments. Instead of retrieving entire documents, the system retrieves smaller segments that contain specific procedural or policy information. The chunking process is defined as:

$$C = \{c_1, c_2, \ldots, c_k\} = \text{Chunk}(\hat{T}) \tag{3}$$

where $\hat{T}$ represents the normalized document text, C denotes the set of semantic chunks, and c_i represents an individual knowledge fragment. Each chunk is associated with metadata attributes describing the source document, structural position, and temporal validity. Metadata enrichment supports contextual filtering and improves retrieval accuracy.

3.5 Semantic Embedding Representation

Each semantic chunk is transformed into a dense vector representation that captures its contextual meaning. Embeddings enable semantic similarity comparisons between queries and knowledge fragments. The embedding function is defined as:

$$v_i = E(c_i), \quad v_i \in \mathbb{R}^d \tag{4}$$

where $E(\cdot)$ denotes the embedding function, c_i represents a semantic chunk, v_i represents the vector embedding of the chunk, and d denotes the embedding dimensionality. Embedding based representation enables retrieval based on semantic similarity rather than purely lexical matching.

3.6 Vector Knowledge Base Construction

The embeddings and corresponding metadata are stored within a unified vector knowledge repository. The knowledge repository is defined as:

$$V = \{(v_i, c_i, m_i)\}_{i=1}^{N} \tag{5}$$

where v_i represents the embedding vector, c_i represents the knowledge fragment, and m_i represents associated metadata. This repository enables efficient similarity search while maintaining traceability between retrieved fragments and original institutional documents.

3.7 Query Context Enrichment

User queries may contain ambiguous terms or context dependent information. Therefore, the system performs query preprocessing and context enrichment before retrieval. The enriched query representation is defined as:

$$\hat{q}_t = f_{\text{context}}(q_t) \tag{6}$$

where q_t represents the user query at time t, $f_{\text{context}}(\cdot)$ represents the query preprocessing function, and $\hat{q}_t$ denotes the enriched query used for retrieval. Context enrichment incorporates conversational history and normalization to improve query interpretation.

3.8 Dense Semantic Retrieval

The enriched query is transformed into an embedding representation and compared with stored knowledge embeddings using cosine similarity. The dense retrieval process is defined as:

$$R_d = \arg\max_{c_i \in V} \cos\left(E(\hat{q}_t), v_i\right) \tag{7}$$

where $E(\hat{q}_t)$ represents the embedding of the query, v_i represents the embedding of knowledge fragment c_i, and $\cos(\cdot)$ denotes cosine similarity. The system retrieves the top k fragments with the highest similarity scores.

3.9 Keyword Based Retrieval

In addition to semantic retrieval, the framework performs keyword based retrieval to capture exact matches for domain specific terminology. This process is defined as:

$$R_k = \text{KeywordRetrieve}(\hat{q}_t, D) \tag{8}$$

where D represents the document corpus and R_k represents the keyword retrieval results. Keyword based retrieval ensures accurate retrieval of structured entities such as policy identifiers, course codes, or procedural terms.

3.10 Hybrid Retrieval and Ranking

The results from semantic retrieval and keyword based retrieval are combined into a unified candidate set:

$$R = R_d \cup R_k \tag{9}$$

The candidate fragments are then ranked using a relevance scoring function:

$$R_t = \text{Rank}(R) \tag{10}$$

where R represents the merged retrieval results, $\text{Rank}(\cdot)$ denotes the ranking function, and R_t represents the final ranked retrieval set. Hybrid retrieval improves both semantic coverage and lexical precision.

3.11 Retrieval Augmented Response Generation

The retrieved knowledge fragments are incorporated into a structured prompt for response generation. The language model generates a candidate response conditioned on retrieved knowledge:

$$G_t = f_{\text{llm}}\left(\langle \hat{q}_t, R_t \rangle\right) \tag{11}$$

where $f_{\text{llm}}(\cdot)$ represents the language model, $\langle \hat{q}_t, R_t \rangle$ denotes the grounded prompt, and G_t represents the generated response. This retrieval augmented generation process reduces reliance on parametric memory and improves factual grounding.

3.12 Rule Based Validation

To ensure policy compliance, the generated response is validated against institutional constraints using rule based verification:

$$G'_t = f_{\text{rule}}(G_t) \tag{12}$$

where G_t represents the generated response, $f_{\text{rule}}(\cdot)$ applies domain specific validation rules, and G'_t represents the validated response. Rule based validation ensures that responses remain consistent with institutional policies.

3.13 Neuro Symbolic Reasoning

Following rule based validation, a neuro symbolic reasoning mechanism verifies logical consistency between the generated response and retrieved knowledge. The reasoning process is defined as:

$$O_t = f_{\text{neuro}}(G'_t) \tag{13}$$

where $f_{\text{neuro}}(\cdot)$ represents the reasoning function and O_t represents the final validated conversational response. Neuro symbolic reasoning combines neural language generation with symbolic constraints to improve response reliability.

3.14 End to End Conversational Workflow

The complete system operates as a sequential composition of all modules described above. The end to end workflow can be summarized as:

$$O_t = f_{\text{neuro}}\big(f_{\text{rule}}\big(f_{\text{llm}}(\langle \hat{q}_t, R_t \rangle)\big)\big) \tag{14}$$

This formulation captures the complete conversational pipeline, integrating knowledge acquisition, retrieval, generation, and validation within a unified framework. By grounding conversational responses in retrieved institutional knowledge and enforcing validation constraints, the proposed framework provides a reliable foundation for conversational academic guidance systems.

4 Methodology

This section describes the methodological procedures used to implement the proposed knowledge centric conversational framework. While Section 3 presented the architectural design of the system, this section focuses on the operational processes that enable knowledge ingestion, semantic representation, hybrid retrieval, response generation, and validation. The methodology follows a knowledge lifecycle perspective in which institutional information is acquired, processed, indexed, retrieved, and validated to support reliable conversational responses. Each stage contributes to improving the contextual relevance, traceability, and reliability of the system.

4.1 Knowledge Ingestion and Preprocessing

The knowledge ingestion process collects institutional documents from diverse academic and administrative sources, including regulations, policies, and notices. Both machine readable and scanned documents are processed using direct extraction and OCR, enabling inclusion of legacy records. The extracted text undergoes preprocessing, including noise removal, normalization, and structural correction, to improve data quality. The cleaned text is then segmented into semantic chunks (e.g., paragraphs or policy statements) to enhance retrieval granularity and relevance.

4.2 Semantic Index Construction

The segmented knowledge fragments are transformed into dense vector embeddings using transformer based models, capturing contextual relationships beyond keyword matching. These embeddings, along with metadata such as document source and category, are stored in a vector database. The resulting index combines semantic embeddings, keyword tokens, and metadata, enabling efficient and context aware retrieval.

4.3 Hybrid Retrieval Strategy

The framework employs a hybrid retrieval strategy combining semantic similarity and keyword based matching. User queries are normalized and enriched before retrieval. Semantic retrieval identifies contextually similar fragments, while keyword retrieval preserves domain specific terms. The results are merged and ranked using relevance scoring based on semantic similarity, lexical overlap, and metadata alignment.

4.4 Retrieval Augmented Response Generation

Retrieved knowledge fragments are incorporated into a structured prompt along with the user query to generate responses using a retrieval augmented generation approach. This grounding in authoritative institutional knowledge significantly improves response accuracy and reliability compared to model centric methods.

4.5 Response Validation and Reliability Control

To ensure reliability, generated responses undergo rule based validation for policy compliance, followed by reasoning based verification to check consistency with retrieved evidence. Unsupported or inconsistent outputs are refined or flagged, reducing hallucinations and improving overall system trustworthiness.

4.6 End to End Conversational Workflow

The methodology operates as an integrated pipeline connecting document ingestion, semantic indexing, retrieval, response generation, and validation. In the offline stage, institutional documents are processed and indexed into structured semantic representations. In the online stage, user queries are handled through hybrid retrieval and retrieval augmented generation to produce grounded responses. A validation module

ensures consistency with institutional knowledge, enabling reliable, transparent, and traceable conversational access.

5 Experimental Setup

This section describes the experimental setup used to evaluate the proposed conversational framework. The experiments assess the effectiveness of the lifecycle based architecture in enabling reliable access to heterogeneous institutional knowledge, focusing on knowledge acquisition, retrieval effectiveness, and response reliability. All configurations use the same dataset, preprocessing pipeline, and query set to ensure fair comparison. The queries represent typical academic information needs, including factual, procedural, and policy related questions.

5.1 Evaluation Dataset and Document Corpus

The evaluation dataset consists of heterogeneous institutional documents collected from multiple academic and administrative domains as summarized in Table 2. These documents include structured records, unstructured textual materials, semi structured forms, and scanned or image based notices. The inclusion of diverse document types enables a comprehensive evaluation of the proposed OCR enabled knowledge acquisition pipeline. Each document is processed through normalization and semantic segmentation before being incorporated into the knowledge repository. Semantic chunks are enriched with structural and temporal metadata to support contextual filtering during retrieval.

Table 2 Institutional Document Corpus Used in Experimental Evaluation

Document Type	**Description**	**Data Format**	**Processing Method**	**No. of Documents**
Structured Documents	Academic calendars, course regulations, timetables, tabular records	Digital (Text / Tables)	Direct text parsing	300
Unstructured Documents	Policy documents, guidelines, notices, announcements	Text / PDF	Text extraction and normalization	600
Semi Structured Documents	Forms, circulars, procedural documents	PDF / Mixed layout	Layout aware parsing	200
Scanned / Image Based Documents	Printed notices and legacy records	Images / Scanned PDFs	OCR (Tesseract + preprocessing)	100

Overall, the corpus contains 1,200 institutional documents, enabling realistic evaluation of heterogeneous knowledge acquisition and conversational retrieval.

5.2 Baseline Systems for Comparative Evaluation

To evaluate the effectiveness of the proposed framework, the system is compared with a set of representative baseline configurations that reflect common conversational and information retrieval architectures as summarized in Table 3. These baselines are designed to isolate the contribution of individual architectural components such as semantic retrieval, OCR based knowledge acquisition, retrieval augmented generation, and reasoning mechanisms.

Table 3 Comparison of Baseline Conversational Systems with the Proposed Knowledge Centric Framework

Baseline System	**Retrieval Method**	**Knowledge Grounding**	**Reasoning Support**
LLM without Retrieval	None	Parametric only	No
Keyword Based System	Lexical matching (TF IDF/BM25)	Document level	No
Dense Retrieval Only System	Semantic embeddings	Chunk level	No
FAQ Based Retrieval System	Predefined Q A matching	Document level	No
Sparse+Dense Retrieval (No Generation)	Hybrid (keyword+dense)	Chunk level	No
Dense Retrieval+Prompt Injection	Semantic embeddings	Chunk level	No
RAG without Validation	Dense retrieval+RAG	Chunk level	No
Hybrid Retrieval+RAG (No Reasoning)	Dense+keyword+RAG	Chunk level	No
OCR Based Keyword Retrieval System	OCR+lexical retrieval	Document level	No
OCR+Dense Retrieval System	OCR+semantic embeddings	Chunk level	No
Rule Based Conversational System	Static rules/flows	Explicit rules	Yes
Proposed OCR Enabled Knowledge Centric System	Hybrid (Dense+Keyword)	Chunk level	Yes

This comparison framework enables systematic analysis of how individual components contribute to retrieval effectiveness and response reliability.

5.3 Evaluation Metrics

To comprehensively evaluate conversational system performance, a multidimensional set of quantitative metrics is employed.

- **Accuracy** measures the proportion of responses that correctly answer user queries based on authoritative institutional knowledge.
- **Precision** evaluates retrieval effectiveness by measuring the correctness of retrieved knowledge fragments.
- **Recall** assesses the systems ability to retrieve all relevant information required to answer a query.
- **F1 score** provides a balanced measure combining precision and recall.
- **Context Relevance** evaluates the semantic alignment between the user query, retrieved knowledge, and generated response.
- **Hallucination Rate** quantifies the proportion of responses containing unsupported or non grounded information.
- **Response Latency** measures the average time required to process a query and generate a response, reflecting system efficiency.

These metrics collectively capture both retrieval quality and conversational reliability.

5.4 Implementation Details

The system is implemented using a modular architecture comprising knowledge ingestion, semantic indexing, hybrid retrieval, and response generation components. OCR based document processing is performed using a standard OCR engine with preprocessing techniques to improve extraction accuracy. Semantic embeddings are generated using transformer based models and stored in a vector database optimized for similarity search. Keyword based retrieval is implemented using an inverted index for efficient lexical matching. Hybrid retrieval integrates semantic similarity and keyword relevance through a rank fusion strategy. Retrieved knowledge fragments are incorporated into structured prompts within a retrieval augmented generation pipeline. All experiments are conducted under consistent hardware and configuration settings to ensure reproducibility and fair comparison.

5.5 Experimental Procedure

The evaluation follows a staged procedure to analyze the contribution of individual components within the knowledge lifecycle. First, the knowledge acquisition module is evaluated to measure OCR extraction accuracy across document types. Second, the knowledge representation module is assessed to study the impact of chunk granularity and metadata enrichment on retrieval performance. Third, the retrieval module is evaluated by comparing lexical, dense, and hybrid retrieval strategies. Fourth, the response generation module is analyzed to examine the effectiveness of retrieval augmented generation in reducing hallucinations. Finally, the complete end to end conversational

framework is evaluated against baseline systems using the defined metrics. This structured evaluation enables a detailed understanding of how each component contributes to overall system performance.

6 Results and Discussion

This section presents the experimental results obtained from evaluating the proposed knowledge centric conversational framework. The analysis focuses on four major aspects of system performance: knowledge acquisition quality, retrieval effectiveness, response generation reliability, and overall system efficiency. The results are compared against baseline conversational architectures to demonstrate the advantages of integrating OCR enabled knowledge acquisition, hybrid retrieval, retrieval augmented generation, and reasoning based validation within a unified framework. The experimental findings are organized to examine how each component contributes to overall system behaviour and how lifecycle integration improves conversational reliability and knowledge grounding.

6.1 Evaluation of the Knowledge Acquisition Module

The Knowledge Acquisition Module is evaluated for its effectiveness in converting heterogeneous institutional documents into accurate textual knowledge. Since OCR quality directly impacts downstream retrieval and response generation, the evaluation considers accuracy, error rates, noise robustness, and processing efficiency. As shown in Table 4, scanned PDFs and printed circulars achieve the highest OCR accuracy (above 93%) with low error rates due to clean layouts, while image based notices exhibit slightly reduced performance because of noise and lighting variations. Handwritten forms and low resolution scans show comparatively higher error rates due to irregular structures and image degradation. Overall, the module demonstrates robust performance across diverse document types, delivering high quality extraction for structured inputs and maintaining reasonable effectiveness for noisy and unstructured documents.

Table 4 Comprehensive OCR and Text Extraction Performance Evaluation Across Document Types

Document Type	**Avg. Pages**	**OCR Accuracy (%)**	**Character Error Rate**	**Word Error Rate**	**Text Completeness**	**Noise Sensitivity**	**Processing Time (ms/page)**
Scanned PDFs	12	93.4	4.8	6.6	High	Moderate	410
Image Based Notices	4	91.2	6.9	8.8	Moderate to High	High	390
Printed Circulars	8	94.1	4.2	5.9	High	Low	370
Handwritten Forms	2	86.7	11.5	14.2	Moderate	Very High	460
Low Resolution Scans	10	88.9	9.3	11.1	Moderate	Very High	445

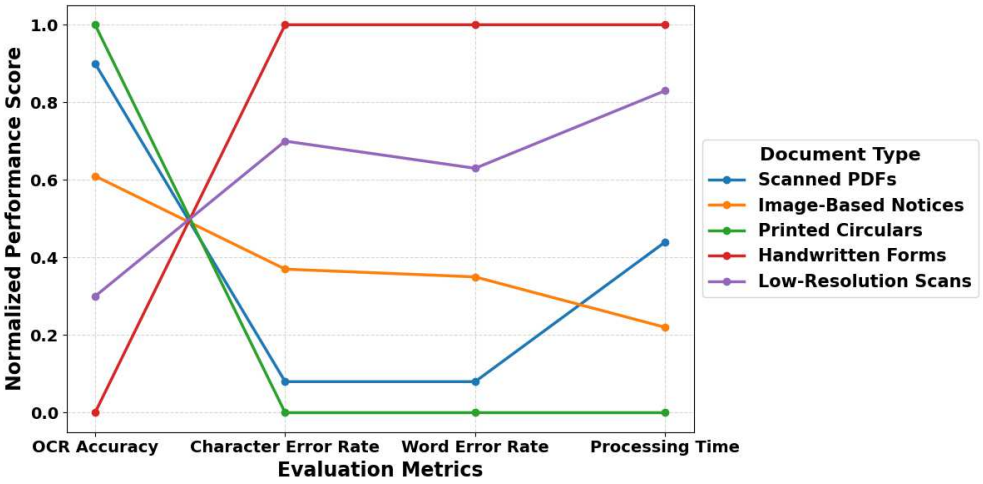

Fig. 2 Parallel Coordinates Visualization of OCR and Text Extraction Performance across document types, showing normalized accuracy, error rates, and processing time for comparative analysis.

Figure 2 presents a parallel coordinates visualization of OCR and text extraction performance across different document types. Unlike conventional bar or pie charts, this representation enables simultaneous analysis of multiple quality and efficiency metrics, revealing performance trade offs for heterogeneous document sources. The results indicate that printed circulars and scanned PDFs exhibit consistently strong OCR performance, whereas handwritten forms and low resolution scans introduce higher error rates and processing overhead. These findings highlight the importance of robust preprocessing mechanisms for noisy institutional documents.

6.2 Evaluation of the Knowledge Representation and Storage Module

The Knowledge Representation and Storage Module is evaluated to assess how different representation strategies affect retrieval readiness within the vector knowledge base. The evaluation focuses on chunk granularity, metadata enrichment, and their influence on semantic retrieval effectiveness. As shown in Table 5, document level storage exhibits the lowest retrieval performance due to coarse grained representations. Chunk level storage significantly improves recall, precision, and context relevance by enabling finer semantic matching. The inclusion of structural and temporal metadata further enhances retrieval effectiveness, with the fully enriched chunk level representation achieving the best overall performance. Although metadata enriched strategies increase index size, the resulting gains in retrieval accuracy and contextual relevance justify the additional storage overhead.

Table 5 Effect of Knowledge Representation Strategies on Retrieval Readiness

Representation Strategy	Chunk Size	Metadata Used	Embedding Dim	Index Size (MB)	Recall	Precision	Context Relevance
Document Level Storage	1200	No	768	180	0.58	0.61	0.55
Chunk Level Storage	300	No	768	245	0.69	0.69	0.69
Chunk + Structural Metadata	300	Yes	768	270	0.72	0.72	0.71
Chunk + Temporal Metadata	300	Yes	768	278	0.73	0.73	0.72
Chunk + Full Metadata	300	Yes	768	290	0.75	0.75	0.73

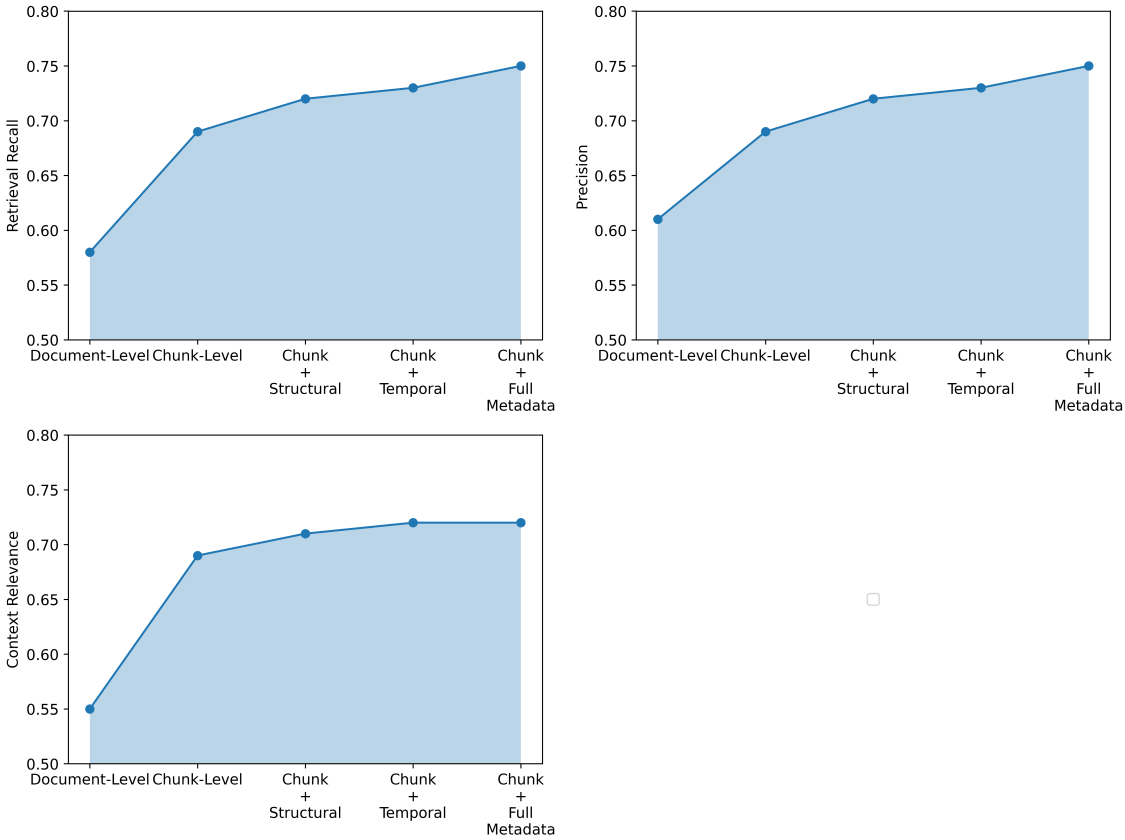

Fig. 3 Knowledge Representation Strategies and Retrieval Readiness across system configurations, illustrating indexing efficiency and retrieval performance trends.

Figure 3 illustrates the impact of knowledge representation strategies on retrieval readiness showing recall, precision, and context relevance across different configurations. As depicted in performance consistently improves from document level storage to chunk based representations, with further gains from structural, temporal, and full metadata integration. The individual area curves highlight that finer grained chunking combined with enriched metadata enables more accurate, precise, and contextually relevant retrieval. The steady upward trend indicates improved semantic alignment and retrieval stability, demonstrating the effectiveness of metadata enhanced representation strategies in supporting high quality conversational retrieval.

6.3 Evaluation of the Knowledge Retrieval Module

The Knowledge Retrieval Module is evaluated to analyze the effectiveness of hybrid strategies combining keyword based and dense semantic retrieval. Retrieval quality is assessed using Precision@k, Recall@k, MRR, NDCG, F1 score, and average latency. As shown in Table 6, keyword based retrieval offers low latency but limited effectiveness, while dense retrieval improves semantic coverage at higher response time. The sparse dense union further enhances performance by leveraging complementary signals, and hybrid rank fusion achieves the best overall results across all quality metrics with acceptable latency. These results highlight a clear performance latency trade off, with hybrid retrieval providing the most balanced solution for knowledge intensive applications.

Table 6 Retrieval Strategy Comparison and Effectiveness Evaluation

Retrieval Strategy	**Top k**	**Precision@k**	**Recall@k**	**MRR**	**NDCG**	**F1 score**	**Avg. Time (ms)**
Keyword Based	5	0.60	0.55	0.49	0.52	0.57	35
Dense Retrieval	5	0.71	0.69	0.63	0.66	0.70	85
Sparse + Dense (Union)	5	0.78	0.76	0.69	0.71	0.77	110
Hybrid (Rank Fusion)	5	0.83	0.81	0.75	0.78	0.82	125

Figure 4 presents a grouped polar bar chart comparing retrieval strategies keyword based, dense retrieval, sparse dense union, and hybrid rank fusion across metrics such as Precision@k, Recall@k, MRR, NDCG, and F1 score. Keyword based retrieval shows the lowest performance, while dense retrieval improves semantic quality. The sparse dense union further enhances results by combining lexical and semantic signals, and hybrid rank fusion achieves the best overall performance. Overall, the figure demonstrates a clear stepwise improvement in retrieval quality from basic keyword methods to hybrid approaches for knowledge intensive applications.

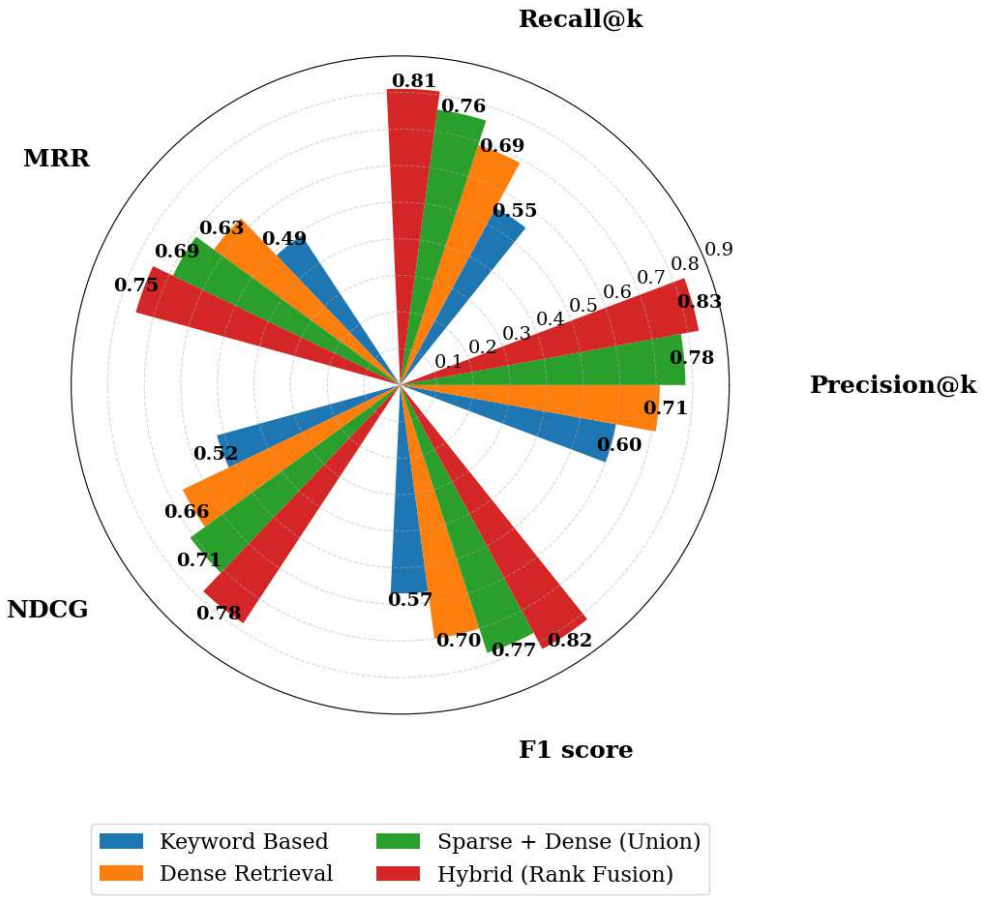

Fig. 4 Retrieval Performance and Latency Trade off across retrieval strategies under varying system configurations and evaluation conditions.

6.4 Evaluation of the Response Generation Module

The Response Generation Module is evaluated to assess response accuracy, contextual grounding, and hallucination behavior under different generation configurations. As shown in Table 7, the LLM only setup exhibits limited grounding and the highest hallucination rate. Incorporating retrieval improves response accuracy and reduces hallucinations, while the proposed retrieval augmented generation pipeline with validated context achieves the best overall performance. These results demonstrate that controlled retrieval grounding significantly enhances response reliability with minimal overhead.

Table 7 Evaluation of Response Generation Quality under Different Configurations

Generation Strategy	Grounding Source	Accuracy (%)	Context Relevance	Faithfulness	Hallucination Rate (%)	Avg. Tokens Generated
LLM Only	None	68.2	0.58	0.55	22.5	180
Dense Retrieval + LLM	Retrieved Chunks	78.4	0.71	0.68	10.9	195
RAG without Validation	Retrieved Chunks	82.9	0.75	0.73	7.6	205
Proposed RAG Pipeline	**Validated Context**	**89.6**	**0.82**	**0.81**	**4.1**	**210**

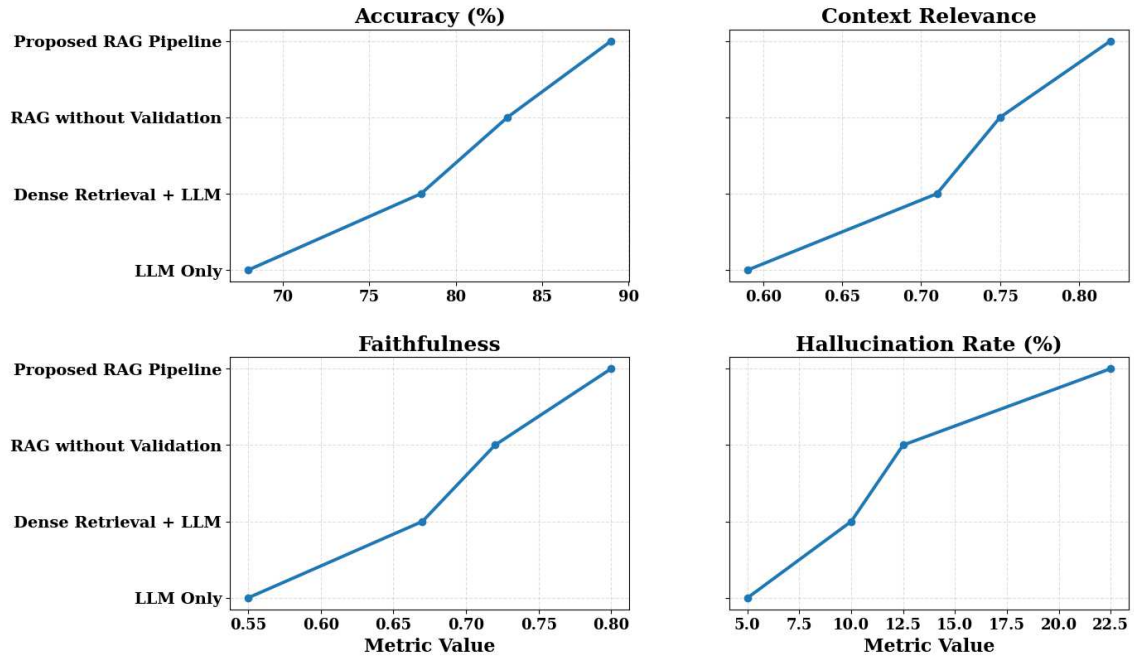

Fig. 5 Accuracy and hallucination comparison of response generation strategies across system configurations, highlighting reliability, consistency, and performance trends.

Figure 5 presents a small multiples analysis of response generation quality across different configurations. The results show consistent improvements in accuracy, context relevance, and faithfulness with the introduction of retrieval grounding and validation, along with a notable reduction in hallucination rate. The proposed retrieval augmented generation pipeline achieves the best overall performance, demonstrating

the effectiveness of validated contextual grounding in enhancing response reliability and consistency.

6.5 Evaluation of the Reasoning and Validation Module

The Reasoning and Validation Module is evaluated for its impact on factual correctness, consistency, and policy compliance. As shown in Table 8, the absence of validation leads to lower accuracy and higher hallucination rates, while rule based validation improves reliability. The addition of neuro symbolic reasoning further enhances factual grounding, with the combined approach achieving the highest accuracy and lowest hallucination rate. These results confirm the effectiveness of integrating symbolic constraints with learned reasoning mechanisms.

Table 8 Effect of Reasoning and Validation on System Performance

Validation Strategy	Policy Rules	Logical Checks	Neuro Symbolic Reasoning	Accuracy (%)	Hallucination Rate (%)	Response Rejection (%)
No Validation	No	No	No	82.9	7.6	0
Rule Based Only	Yes	Yes	No	86.7	5.3	6.2
Neuro Symbolic Only	No	Yes	Yes	87.9	4.8	7.1
Rule + Neuro Symbolic	**Yes**	**Yes**	**Yes**	**89.6**	**4.1**	**8.5**

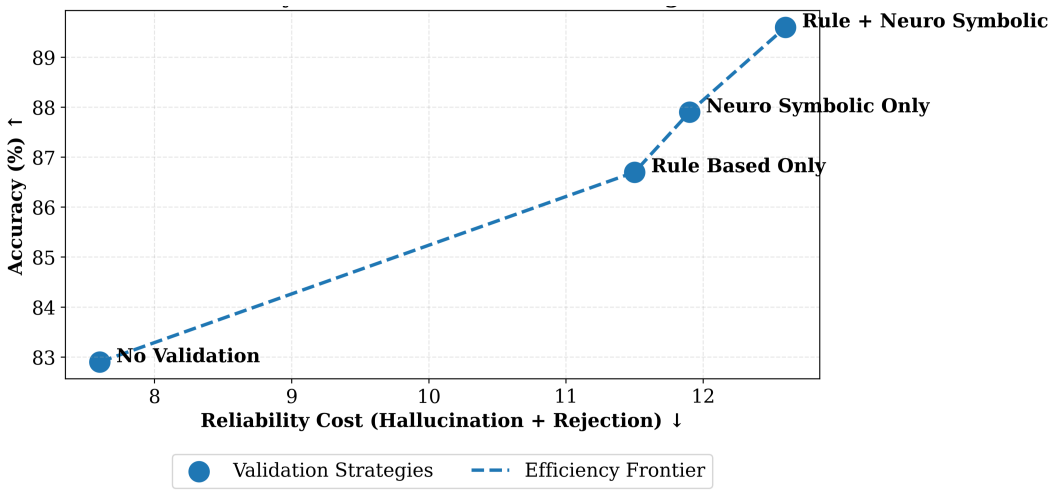

Fig. 6 Effect of reasoning and validation on response quality across system configurations, emphasizing accuracy, reliability, and consistency improvements.

Figure 6 illustrates the impact of different reasoning and validation strategies on response quality. The introduction of rule based and neuro symbolic validation leads

to a steady increase in accuracy and a corresponding reduction in hallucination rate compared to no validation. The combined rule + neuro symbolic approach achieves the best overall performance, demonstrating improved reliability and consistency, while also resulting in a higher response rejection rate due to stricter validation constraints.

6.6 Evaluation of System Efficiency and Overall Performance

The end to end system is evaluated in terms of latency and resource efficiency under real time conditions. As shown in Table 9, keyword based retrieval achieves the lowest latency, while dense retrieval and retrieval augmented generation incur higher response times due to semantic processing. Although the proposed system introduces validation overhead, latency remains within acceptable limits for interactive use, demonstrating that improved reliability can be achieved with a manageable trade off in performance and resource consumption.

Table 9 System Efficiency and End to End Performance Evaluation

System Configuration	Avg. Latency (ms)	Retrieval Time (ms)	Generation Time (ms)	Validation Time (ms)	Memory Usage (MB)
Keyword Based System	180	35	120	0	620
Dense Retrieval Only	260	85	150	0	740
RAG without Validation	290	125	150	0	810
Proposed System	**310**	**125**	**150**	**35**	**860**

Figure 7 presents a multidimensional performance analysis using Andrews curves to compare system configurations across efficiency and performance metrics. The proposed system exhibits a more complex efficiency profile due to the inclusion of validation and reasoning mechanisms, which introduce additional computational overhead and memory usage. Despite this added complexity, the system maintains retrieval and generation performance comparable to other retrieval augmented configurations, indicating that validation overhead does not significantly impact core operational efficiency. This balance demonstrates that enhanced reliability and validation can be achieved without compromising overall system responsiveness.

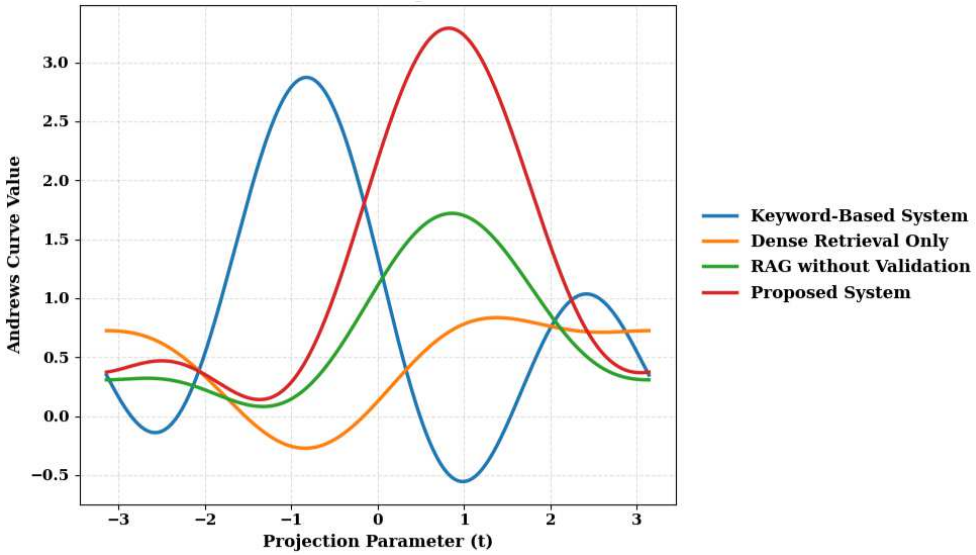

Fig. 7 Multidimensional performance analysis using Andrews curves for comparative evaluation across system configurations and performance metrics.

6.7 Overall Performance Comparison

This section compares the proposed knowledge centric conversational system with baseline and retrieval based approaches using response quality and efficiency metrics. As shown in Table 10, systems without retrieval suffer from low accuracy, weak grounding, and high hallucination rates. Keyword based and dense retrieval approaches provide limited improvements in latency and semantic coverage, while unvalidated retrieval augmented generation enhances precision and recall but retains residual hallucinations. The proposed system outperforms all baseline configurations across core quality metrics while maintaining acceptable latency, demonstrating the effectiveness of integrating OCR based knowledge acquisition, hybrid retrieval, and validation mechanisms within a unified framework.

Figure 8 presents a multi metric radar chart comparing conversational systems across accuracy, precision, recall, F1 score, contextual relevance, anti hallucination capability, and efficiency. Metrics are normalized for uniform comparison, with larger radial distance indicating better performance, and inverse normalization applied for hallucination. The proposed knowledge centric system forms the most balanced and outermost polygon, demonstrating superior performance across all quality dimensions while maintaining competitive efficiency. In contrast, keyword based retrieval achieves high efficiency but lower semantic effectiveness, and the standalone LLM exhibits weak grounding due to higher hallucination. The progressive expansion of the radar area highlights the cumulative benefits of hybrid retrieval, validation, and reasoning.

Table 10 Overall Performance Comparison Across Conversational Systems

Metric / System	LLM	Keyword	Dense	RAG	Hybrid	OCR+Dense	Proposed
Accuracy (%)	68.2	61.4	73.6	82.9	85.6	74.8	**89.6**
Precision	0.65	0.60	0.71	0.80	0.83	0.72	**0.86**
Recall	0.62	0.55	0.69	0.78	0.81	0.70	**0.84**
F1 score	0.63	0.57	0.70	0.79	0.82	0.71	**0.85**
Context Relevance	0.58	0.52	0.66	0.75	0.78	0.67	**0.82**
Hallucination Rate (%)	22.5	18.3	12.1	7.6	6.2	11.5	**4.1**
Latency (ms)	420	180	260	290	300	265	**310**

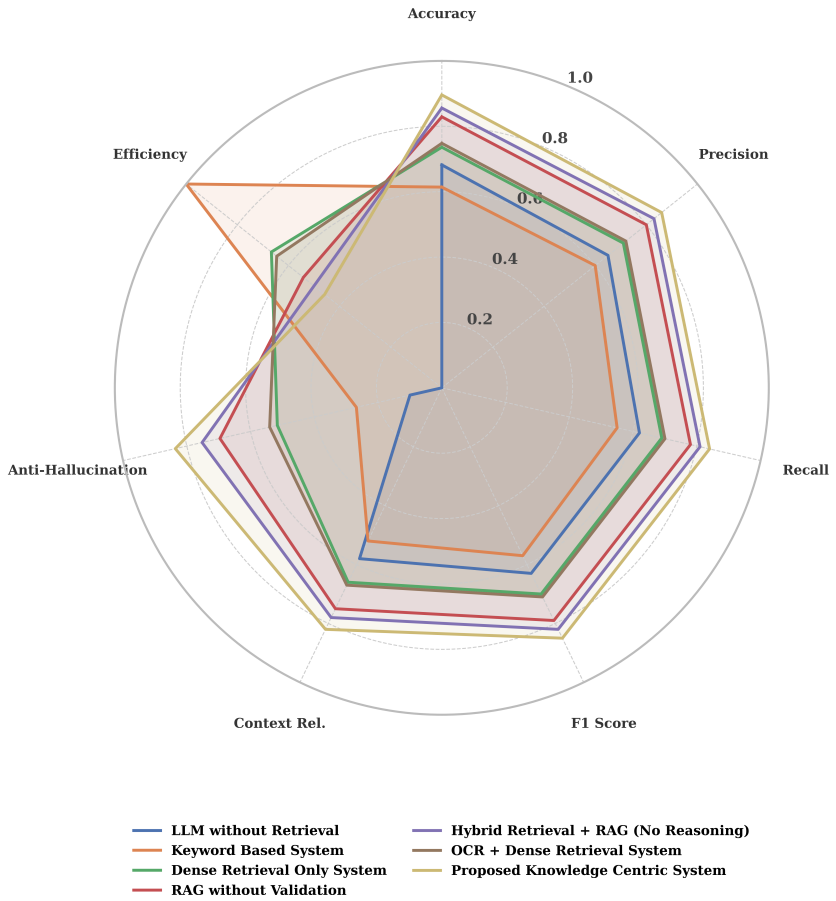

Fig. 8 Multi metric performance comparison across conversational systems under diverse experimental configurations and standardized evaluation benchmarks.

6.8 Ablation Study: Component Contribution Analysis

An ablation study evaluates the contribution of individual components within the proposed knowledge centric conversational framework by selectively disabling key modules. Removing OCR reduces coverage of legacy and scanned documents, while replacing semantic embeddings with keyword based indexing weakens semantic alignment, particularly for paraphrased queries. Eliminating hybrid retrieval lowers recall, and disabling retrieval augmented generation significantly increases hallucination rates. Further, removing reasoning and validation leads to higher policy violations and reduced reliability. As shown in Table 11, performance consistently degrades when components are removed, confirming that the effectiveness of the proposed system arises from the synergistic integration of OCR, hybrid retrieval, RAG, and validation mechanisms.

Table 11 Impact of System Components on Accuracy and Hallucination Rate

System Configuration	Active Components	Accuracy (%)	Hallucination Rate (%)
Proposed System	OCR, Hybrid Retrieval, RAG, Reasoning	**89.6**	**4.1**
Without Reasoning	OCR, Hybrid Retrieval, RAG	85.6	6.2
Without RAG	OCR, Hybrid Retrieval, LLM	78.4	10.9
Dense Retrieval Only	OCR, Dense Retrieval, LLM	82.1	7.8
Keyword Retrieval Only	OCR, Keyword Retrieval, LLM	74.3	12.6
Without OCR	Text Corpus, Hybrid Retrieval, RAG, Reasoning	80.7	8.9
LLM Baseline	LLM Only	68.2	22.5

7 Conclusion and Future Work

This paper presented a knowledge centric conversational framework for reliable access to heterogeneous institutional knowledge. The proposed approach integrates OCR enabled knowledge acquisition, semantic representation through embedding based indexing, hybrid retrieval, retrieval augmented generation, and reasoning driven validation within a unified architecture. By grounding responses in institutional documents and enforcing validation constraints, the framework addresses key limitations of model centric systems, including hallucinations and weak contextual grounding.

Experimental results demonstrate that each component contributes to improved performance. OCR enhances knowledge coverage, semantic chunking and metadata improve retrieval quality, and hybrid retrieval increases recall and ranking accuracy. The integration of retrieval augmented generation with validation significantly reduces hallucinations and improves factual consistency. Comparative evaluation shows that the proposed system outperforms baseline methods across accuracy, precision, recall, F1 score, context relevance, and hallucination rate, while maintaining acceptable latency for real time interaction.

The framework is adaptable to other document intensive domains such as enterprise knowledge management, healthcare, legal advisory, and technical support, owing to its modular design.

Future work will focus on dynamic knowledge base updates, multilingual and cross lingual query support, and explainable reasoning for improved transparency. Additionally, large scale user centric evaluations will be conducted to assess real world usability. Overall, this work demonstrates that integrating knowledge acquisition, hybrid retrieval, retrieval augmented generation, and validation provides a scalable and reliable foundation for trustworthy conversational systems.

Data Availability

The institutional documents used in this study are not publicly available due to confidentiality constraints of the academic institution. Access to the dataset may be granted by the corresponding author upon reasonable request. The source code and implementation of the proposed system are openly available at https://doi.org/10.5281/zenodo.19230595 and https://github.com/Kesinenibharath/College_Chatbot/tree/v1.0.0.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank Vignan's Lara Institute Of Technology and Science for providing the necessary infrastructure and support to carry out this research.

Funding

No external funding was received for this study.

References

[1] Kacupaj E, Singh K, Maleshkova M, Lehmann J (2024) Conversational question answering over knowledge graphs. Springer. https://doi.org/10.1007/978-3-031-64451-1_9

[2] Rezaei M (2025) Artificial intelligence in knowledge management: Identifying and addressing the key implementation challenges https://doi.org/10.1016/j.techfore.2025.124183

[3] Remadi A, Hage K El, Hobeika Y, Bugiotti F (2024) To prompt or not to prompt: Navigating the use of Large Language Models for integrating and modeling heterogeneous data. https://doi.org/10.1016/j.datak.2024.102313

[4] Atzeni P, Baldazzi T, Bellomarini L, Laurenza E, Sallinger E (2026) Semantic-aware query answering with large language models. Data Knowl Eng. https://doi.org/10.1016/j.datak.2025.102494

[5] Bashir A, Peng R, Ding Y (2025) Logic-infused knowledge graph QA: Enhancing large language models for specialized domains through Prolog integration. Data Knowl Eng. https://doi.org/10.1016/j.datak.2025.102406

[6] Lewis P, Perez E, Piktus A,Petroni F,Karpukhin V,Goyal N, Kuttler H,Lewis M (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS. https://doi.org/10.48550/arXiv.2005.11401

[7] Simsek U, Karle E, Angele K, Huaman E, Opdenplatz J, Sommer D, Umbrich J, Fensel D (2023) A knowledge graph perspective on knowledge engineering. SN Comput Sci. https://doi.org/10.1007/s42979-022-01429-x

[8] Peng C, Xia F, Naseriparsa M, Osborne F (2023) Knowledge graphs: Opportunities and challenges. Artif Intell Rev. https://doi.org/10.1007/s10462-023-10465-9

[9] Sharma A, Kumar S (2023) Ontology-based semantic retrieval of documents using Word2vec model. Data Knowl Eng. https://doi.org/10.1016/j.datak.2022.102110

[10] Touvron H, Lavril T, Martinet X, Lacroix T, Roziere B, Goyal N, Hambro E, Azhar F (2023) LLaMA: Open and efficient foundation language models. arXiv. https://doi.org/10.48550/arXiv.2302.13971

[11] Song M (2026) Defining the problem: The impact of OCR quality on retrieval-augmented generation performance and strategies for improvement. https://doi.org/10.1016/j.ipm.2025.104368

[12] Li Z, Wang Z, Wang W, Hung K, Xie H, Wang F (2025) Retrieval-augmented generation for educational application: A systematic survey. Comput Educ AI. https://doi.org/10.1016/j.caeai.2025.100417

[13] Liu H, Yang S, Shi G, Miao Z (2025) Knowledge graph reasoning: Mainstream methods, applications and prospects. Eng Appl AI. https://doi.org/10.1016/j.engappai.2025.111625

[14] Liu X, Mao T, Shi Y, Ren Y (2024) Overview of knowledge reasoning for knowledge graphs. Neurocomputing. https://doi.org/10.1016/j.neucom.2024.127571

[15] Tian L, Zhou X, Wu Y, Zhou W, Zhang J, Zhang T (2022) Knowledge graph and knowledge reasoning: A systematic review. J Electron Sci Technol. https://doi.org/10.1016/j.jnlest.2022.100159

[16] Ghali M, Farrag A, Won D, Jin Y (2025) Enhancing knowledge retrieval with in-context learning and semantic search through generative AI. Knowl Based Syst.

https://doi.org/10.1016/j.knosys.2025.113047

[17] Siddharth L, Luo J (2024) Retrieval augmented generation using engineering design knowledge. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2024.112410

[18] Su J, Zhou J, Zhang Z, Nakov P, Cardie C, Peng Zhou J (2025) Towards More Robust Retrieval-Augmented Generation: Evaluating RAG Under Adversarial Poisoning Attacks. arXiv. https://doi.org/10.48550/arXiv.2412.16708

[19] Linders J, Tomczak JM (2025) Knowledge graph-extended retrieval augmented generation for question answering. Appl Intell. https://doi.org/10.1007/s10489-025-06885-5

[20] Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, Yang L, Zhang W (2026) Retrieval-Augmented Generation for AI-Generated Content: A Survey. Data Sci Eng. https://doi.org/10.1007/s41019-025-00335-5

[21] Sonawane SS, Mahalle PN, Ghotkar AS (2022) Information Retrieval and Natural Language Processing A graph theory approach. Springer. https://doi.org/10.1007/978-981-16-9995-5

[22] Klesel M, Wittmann H (2025) Retrieval-augmented generation. Bus Inf Syst Eng. https://doi.org/10.1007/s12599-025-00945-3

[23] Wang S, Yang H, Liu W (2025) Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph. Sci Rep. https://doi.org/10.1038/s41598-025-21222-z

[24] Abdelmagied M, Chatti M, Joarder S, Ain Q, Alatrash R (2025) Leveraging Graph Retrieval-Augmented Generation to Support Learners Understanding of Knowledge Concepts in MOOCs. https://doi.org/10.1007/978-3-032-00056-9_10

[25] Diamantini C, Mele A, Mircoli A, Potena D, Rossetti C, Storti E (2025) A Graph RAG Approach to Enhance Explainability in Dataset Discovery. Data Sci Eng. https://doi.org/10.1007/s41019-025-00313-x

[26] Wang Y, Wan Y, Lei X, Chen Q, Hu H (2025) A retrieval augmented generation based optimization approach for medical knowledge understanding and reasoning in large language models. https://doi.org/10.1016/j.array.2025.100504

[27] Rawat R, Borana K, Gehlot S, Rawat A, Rawat H, Singh Dangi C (2026) Information retrieval framework using knowledge graph embeddings and uncertainty modelling using probabilistic soft logic. Discover Comput. https://doi.org/10.1007/s10791-025-09859-w

[28] Wang Y, Lin Q, Shi J, An Y, Liu J, Wei B, Jiang X (2026) Explainable visual question answering: A survey on methods, datasets and evaluation. https://doi.org/10.1016/j.inffus.2026.104215

[29] Chen L, Pardeshi M, Liao Y, Pai K (2025) Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model. Comput Stand Interfaces 94. https://doi.org/10.1016/j.csi.2025.103995

[30] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Sun J, Wang M, Wang H (2023) Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv. https://doi.org/10.48550/arXiv.2312.10997

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AuthorBiographies.pdf
- Highlights.pdf
- GraphicalAbstract.pdf