Journal of Central Banking Theory and Practice, 2026, 2, pp. 99-123
Received: 18 December 2025; accepted: 29 March 2026

UDC: 336.71:004.8(497.6)
DOI: 10.2478/jcbtp-2026-0014

***Kemal Hajdarevic*[*]*, Jasmina Selimovic*[**]**

[*] Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Email: khajdarevic@etf.unsa.ba

[**] Central Bank of Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina

Email: Jasmina.Selimovic@cbbh.ba

Offline NLP Assistants for Internal Knowledge Access in a Central Bank: Proof-of-Concept Systems for Semantic Search and HR Chatbot

Abstract: This paper presents two offline, on-premise NLP proof-of-concept assistants built on a shared architecture for internal knowledge access in the Central Bank of Bosnia and Herzegovina: (i) a semantic document search tool for internal Word/PDF repositories and (ii) an HR chatbot that applies retrieval-augmented generation (RAG) over indexed HR policies and procedures. Rather than proposing a novel NLP method, the paper contributes by documenting a reusable offline architecture for institutional AI assistants in a security-constrained central banking environment and by providing pilot evidence on how established semantic retrieval and RAG techniques can be adapted to strict requirements of confidentiality, data sovereignty, and governance. The semantic search assistant combines exact phrase matching with embedding-based retrieval and hybrid re-ranking, while the HR chatbot generates source-grounded answers using locally hosted language models under explicit governance constraints, including B/H/S-only output, strict fallback behaviour, and transparent display of retrieved passages. Pilot results indicate that hybrid retrieval offers the most reliable performance across representative internal queries, while the HR chatbot demonstrates the feasibility of document-grounded employee support under offline institutional constraints. The findings provide preliminary evidence that offline NLP assistants can improve access to internal institutional knowledge while remaining compatible with the security and operational risk requirements typical of central banking environments.

Keywords: semantic search, document retrieval, sentence embeddings, transformer models, natural language processing, central banking, proof-of-concept.

JEL Code: E52, E58, O33, C88

1. Introduction

Natural Language Processing (NLP) has become an important enabler of intelligent access to unstructured textual information, particularly in organizations that rely heavily on document-based knowledge. Recent advances in transformer models and sentence embeddings have improved the ability of systems to retrieve information on the basis of semantic meaning rather than exact keyword overlap, making such approaches increasingly relevant for professional and regulatory environments.

Central banks represent a domain in which access to internal documents is both critical and challenging. A substantial part of institutional knowledge is stored in unstructured formats such as internal reports, regulatory texts, policy analyses, procedures, and correspondence, most commonly in Word and PDF files. These documents are often distributed across multiple directories and repositories, accumulated over many years, and written by different organizational units. As a result, conventional keyword-based search frequently proves insufficient, especially when users are unfamiliar with the exact terminology used in the source documents.

At the same time, the use of AI-based tools in central banking is constrained by strict requirements related to confidentiality, security, data sovereignty, and operational risk. In many such environments, sensitive internal data cannot be exposed to external cloud services, and reliance on internet-based AI processing is unacceptable. These constraints create demand for offline, on-premise AI systems that can improve access to institutional knowledge while remaining compatible with internal governance and security requirements.

In addition to document retrieval for analytical and supervisory work, similar challenges also arise in internal support functions such as human resources. HR policies, procedures, and benefits guidance are often dispersed across unstructured documents, leading to repetitive employee queries and inconsistent interpretation of internal rules. This creates a suitable use case for document-grounded conversational assistants operating under controlled institutional constraints.

This paper presents two offline, on-premise proof-of-concept (PoC) NLP assistants developed for the Central Bank of Bosnia and Herzegovina: (i) a semantic document search system for internal Word/PDF repositories and (ii) an HR chatbot for employee support based on retrieval-augmented generation (RAG). The purpose of the paper is not to propose a novel retrieval or language-generation method, but to examine how established NLP techniques can be operationally and institutionally deployed in a central bank setting under strict offline security and governance constraints.

The contribution of the paper is threefold. First, it documents a reusable offline architecture for NLP-based assistants suitable for central banking environments characterized by strict confidentiality and operational risk requirements. Second, it shows how established techniques in semantic retrieval, sentence embeddings, and retrieval-augmented generation can be adapted to institutional requirements such as on-premise deployment, source-grounded outputs, controlled fallback behaviour, and restricted language generation. Third, it provides operational insights from pilot implementation of two practical assistants—a semantic search tool and an HR chatbot, highlighting deployment trade-offs, governance considerations, and structured pilot evaluation results relevant for similar regulated public-sector institutions.

Taken together, these contributions are not aimed at algorithmic novelty, but at the design and pilot evaluation of a reusable offline NLP assistant architecture adapted to the governance, confidentiality, and operational constraints of a central bank.

Although the proposed system was developed and evaluated within the Central Bank of Bosnia and Herzegovina, the architectural principles described in this study may also be relevant for other central banks and regulated public-sector organizations operating in similarly restricted IT environments.

2. Literature review: AI-enabled internal knowledge access in central banking

2.1. AI and NLP in central banking

Artificial intelligence (AI) and data-driven analytics are increasingly used in both commercial and central banking. In central banks, much of this work has focused on macroeconomic analysis, financial stability, supervisory technology, regulatory reporting, and broader challenges related to fintech and institutional

adaptation (Araujo et al., 2024; Doerr et al., 2021; Grigorescu, 2024; Vučinić & Luburić, 2024). International initiatives such as the BIS Central Bank Language Models (CB-LMs) further illustrate growing interest in NLP applications for central bank communication and analysis (Araujo et al., 2024; Gambacorta et al., 2024).

However, comparatively less attention has been devoted to practical, user-facing systems that support day-to-day access to internal institutional knowledge. This is an important gap because a substantial share of operational knowledge in central banks remains embedded in unstructured internal documents, including reports, procedures, regulatory materials, and correspondence, most often stored in Word and PDF formats.

2.2. Access to unstructured internal documents

Retrieval from unstructured document repositories is still often based on keyword search. Although useful for known-item lookup, such approaches may perform poorly when users are unfamiliar with the exact wording used in source documents or when terminology differs across departments and time periods. In banking and related institutional settings, prior work suggests that valuable knowledge is frequently dispersed across heterogeneous document collections, limiting efficient reuse and institutional learning (Cimpeanu et al., 2023; Koti, 2024).

These challenges are particularly relevant in central banks, where internal documents are distributed across multiple units and often subject to strict confidentiality and compliance requirements. As a result, there is a practical need for retrieval systems that can improve access to internal knowledge without relying on externally hosted AI services.

2.3. Semantic search and sentence embeddings

Recent advances in NLP have made semantic search increasingly viable through the use of embeddings. Sentence embedding models map text segments into vector spaces in which semantic similarity can be measured computationally, allowing retrieval based on meaning rather than exact lexical overlap. This is particularly useful in professional environments where users may express queries in conceptual rather than document-specific language.

Transformer-based multilingual embedding models are especially relevant for institutional repositories with mixed-language content. By indexing overlapping chunks rather than full documents, semantic search systems can retrieve specific passages that are more directly useful to end users. Hybrid retrieval approaches that combine semantic similarity with lexical matching have been proposed to balance recall and precision, especially in domains where formal terminology still matters.

2.4. Retrieval-augmented generation in regulated environments

Large language models can provide flexible natural-language responses, but their use in regulated institutional environments is constrained by the risk of unsupported or hallucinated outputs. Retrieval-augmented generation (RAG) addresses this limitation by grounding answer generation in retrieved passages from a trusted document base. In this architecture, relevant fragments are first retrieved and then used as contextual input for generation, allowing outputs to remain traceable to source documents.

For central banking environments, RAG is attractive because document ingestion, embedding generation, retrieval, and language-model inference can all be implemented on-premise. This makes it possible to combine conversational access with stronger control over confidentiality, data sovereignty, and governance than is typically feasible in cloud-based deployments.

2.5. On-premise AI, governance, and security constraints

Security, confidentiality, and operational risk are central considerations in AI adoption within central banks. More broadly, prior work on automation in banking has also shown that technology adoption in regulated financial institutions must be aligned with control, governance, and operational-risk considerations (Villar & Khan, 2021). Many institutions operate in restricted network environments where sensitive internal data cannot be transferred to external services. Consequently, AI systems intended for internal knowledge access must often be designed for offline or fully on-premise deployment.

Recent literature emphasizes the importance of trustworthy and governed AI systems in regulated financial institutions, including transparency, bounded outputs, and institutional control over model behaviour (Araujo et al., 2024; Grigorescu, 2024; Soundenkar et al., 2024). In this context, architectural choices such as

local inference, source-grounded outputs, and fallback behaviour are not merely technical preferences, but governance-relevant design decisions.

2.6. HR chatbots and employee self-service

Conversational agents are increasingly used in human resource management as part of broader employee self-service systems. HR chatbots typically provide access to policies, procedures, and routine guidance, helping reduce repetitive inquiries and improve consistency of interpretation (Majumder & Mondal, 2021). Recent research also highlights the importance of document grounding, transparency, and governance when generative AI is applied in HR contexts (Bernik & Šprajc, 2025; Budhwar et al., 2023; Li & Cheng, 2025).

These issues are especially important in regulated institutions, where HR-related responses must remain bounded, policy-based, and auditable. This makes HR a suitable use case for controlled, document-grounded conversational assistants operating entirely within institutional infrastructure.

2.7. Research gap

Although existing literature documents growing use of AI and NLP in central banking, most prior work focuses on analytical applications rather than operational systems for internal knowledge access. At the same time, research on semantic search and RAG is well established in the broader NLP literature, but relatively little has been published on how such methods can be adapted to the confidentiality, governance, and infrastructure constraints typical of central banks.

This gap motivates the present study. Rather than proposing a new retrieval or generation method, the paper examines how established semantic retrieval and RAG techniques can be implemented as offline, on-premise assistants for internal document access in a central banking environment. The contribution, therefore, lies in institutional adaptation, governance-aware system design, and pilot evaluation in a real organizational setting.

3. System architecture and implementation

3.1. Overall Architecture: Two Assistants, One Offline Pipeline

The proposed proof-of-concept system is designed as an on-premise, offline-capable architecture for secure access to internal knowledge resources. Its design reflects the operational constraints of a central bank where document confidentiality, restricted connectivity, and institutional control over AI outputs are primary requirements.

At a high level, the system consists of four modules: (i) document ingestion and preprocessing, (ii) chunking and embedding generation, (iii) retrieval and ranking, and (iv) a web-based user interface. All components execute locally within institutional infrastructure, without reliance on external APIs or cloud services. This architecture ensures data sovereignty while allowing modern NLP techniques to be applied to internal document collections.

A key architectural feature is that the same offline pipeline is reused by two assistants with different user-facing functions. The first is a semantic document search assistant focused on retrieval of relevant document passages. The second is an HR chatbot that builds on the same retrieval pipeline and extends it with controlled language generation through a retrieval-augmented generation (RAG) workflow. This shared architecture reduces implementation complexity and supports consistent governance across both use cases.

3.2. Shared Document Ingestion, Chunking and Indexing Pipeline

The system processes internal document collections stored in Microsoft Word (.docx) and PDF (.pdf) formats. During indexing, the ingestion module recursively scans a user-specified root directory and its subdirectories to identify eligible documents.

Extracted text is normalized to remove excessive whitespace and formatting artefacts. Documents are then segmented into overlapping text chunks, which serve as the basic retrieval units. This chunk-based representation balances two competing objectives: preserving enough local context for meaningful semantic retrieval while keeping the units small enough for efficient indexing and ranking. Overlap between adjacent chunks helps retain information that might otherwise be split across chunk boundaries.

Each chunk is encoded using a multilingual sentence embedding model and stored together with its metadata in a local index. Embedding generation is performed in batches for efficiency, while model-specific caching is used to avoid unnecessary recomputation when indexed documents remain unchanged. This design supports practical incremental indexing and is particularly useful in institutional environments where document repositories evolve gradually rather than continuously.

3.3. Design rationale for offline indexing

The indexing pipeline was designed to support a medium-scale, heterogeneous document corpus under restricted infrastructure conditions. Three design choices are particularly important.

First, multilingual sentence embeddings were selected because the institutional document environment is linguistically mixed and cannot assume English-only content. Second, chunk-level indexing was preferred over full-document indexing because users often need access to specific passages rather than entire files. Third, local caching of embeddings was introduced to reduce repeated indexing costs and make the system operationally sustainable on standard on-premise hardware.

These choices do not represent methodological novelty; rather, they reflect the adaptation of established NLP practices to the practical and governance constraints of a central banking environment.

3.4. Assistant A: Semantic Document Search

The first assistant provides semantic retrieval over internal document repositories. It is intended for use cases in which staff need to locate relevant passages across heterogeneous Word and PDF collections without knowing the exact wording, file location, or document source in advance.

Document chunks are represented in a shared vector space using multilingual sentence embeddings, enabling similarity-based retrieval at query time. Several pre-trained embedding models were considered in order to accommodate different trade-offs between semantic quality, indexing speed, and hardware requirements. All models are used strictly in inference mode, and all processing remains within local institutional infrastructure.

This design is particularly suitable for central banking environments because it supports meaning-based retrieval while preserving confidentiality and avoiding reliance on external services. It also allows model selection to be adapted to operational needs, ranging from lightweight indexing for large collections to higher-capacity models for more semantically demanding queries.

3.4.1. Search Modes and Retrieval Logic

The semantic search assistant supports three complementary retrieval modes.

The first mode is exact phrase search, intended for cases where precise wording matters, such as references to formal terms, legal provisions, or procedural expressions. The second mode is semantic similarity search, where the user query is embedded into the same vector space as indexed chunks and compared by vector similarity. This mode supports concept-based retrieval even when the user's wording differs from that used in the documents.

The third mode is a hybrid approach that combines semantic retrieval with lexical overlap signals through re-ranking. This mode is particularly useful in institutional document collections, where both conceptual relevance and terminological precision may matter. Optional filters by file type and directory path further allow narrowing of the search space when needed.

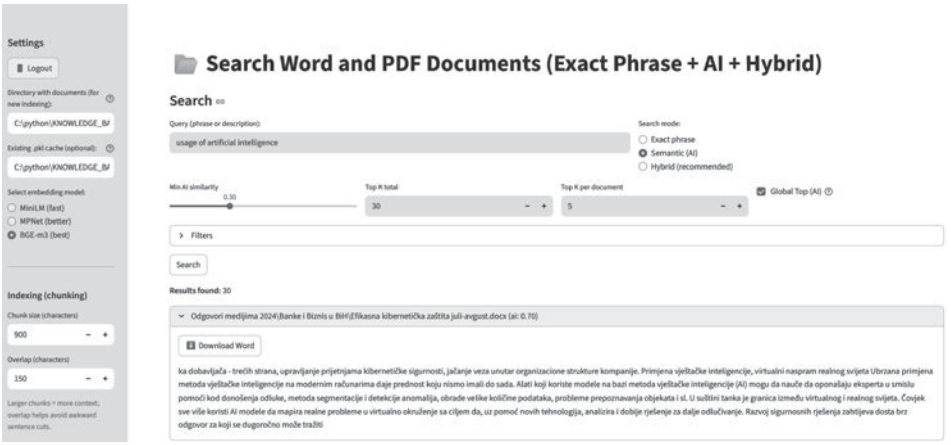
The three modes were included not to introduce a new retrieval method, but to reflect different categories of information needs observed in professional document use: known-item lookup, exploratory search, and mixed analytical queries.

3.4.2. User Interface and Interaction

Both assistants are accessed through a lightweight web-based interface. The interface supports authenticated use, configuration of indexing and retrieval settings, and transparent inspection of returned results.

For the semantic search assistant, retrieved fragments are displayed together with relevance indicators and source references. This presentation is intended to support user verification and interpretability, rather than treating the system as a fully autonomous decision tool. Figure 1 illustrates the interface of the semantic document search assistant.

Figure 1: User interface of the NLP-driven semantic document search system with hybrid retrieval results



3.5. Assistant B: HR Chatbot (RAG Workflow)

The second assistant is an HR chatbot that reuses the same offline ingestion and embedding pipeline as the semantic search system but extends it with retrieval-augmented generation using a locally hosted LLM.

The purpose of the HR chatbot is not to provide open-ended conversational intelligence, but to offer concise, policy-grounded assistance for routine employee questions based on authoritative internal HR documents. This use case was selected because HR queries are often repetitive, document-based, and suitable for constrained answer generation under strong governance requirements.

3.5.1. Retrieval stage

In the HR chatbot, retrieval is the first controlled stage of the RAG workflow. HR policies, procedures, and internal guidelines are segmented into semantically coherent chunks and indexed in a local FAISS vector store using multilingual sentence embeddings. At query time, the system retrieves the most relevant passages based on vector similarity.

To limit noise and unnecessary context, retrieval is constrained to TOP_K = 3. This setting is appropriate for HR content, where concise and authoritative excerpts are generally preferable to a larger set of partially relevant passages. Chunk size and overlap are aligned with the structure of HR documents so that retrieved fragments typically preserve complete policy statements or procedural instructions.

3.5.2. Prompt template and response policy

Answer generation is governed by a fixed prompt template that enforces institutional constraints and consistent system behaviour. The model is instructed to respond exclusively in Bosnian/Croatian/Serbian (B/H/S) and to use only the passages retrieved during the retrieval stage. It is explicitly prohibited from relying on unsupported assumptions or external knowledge.

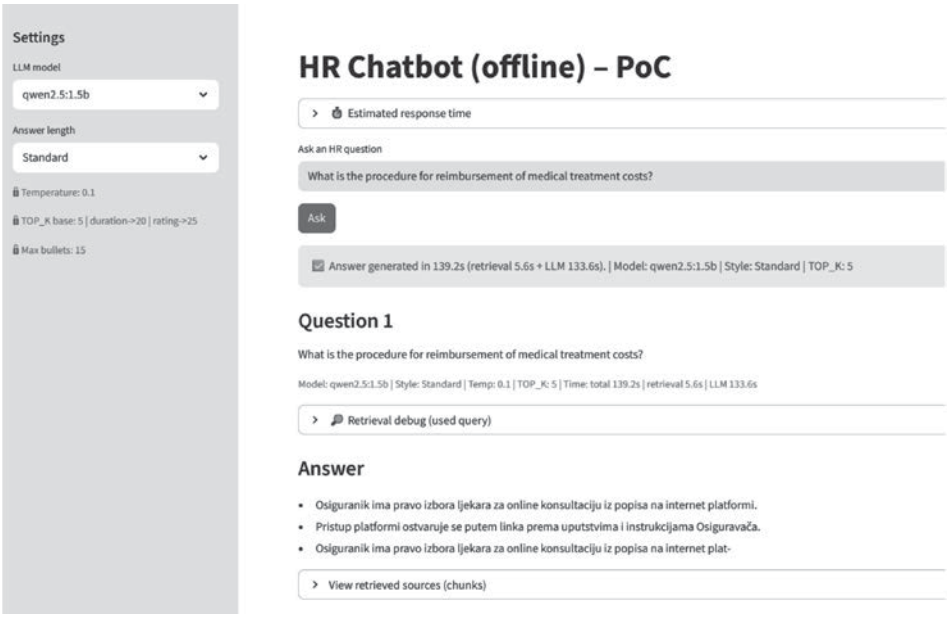
Responses are limited to a maximum of six bullet points in order to promote concise, policy-oriented answers and predictable offline inference time. When the retrieved context is insufficient to support a reliable answer, the chatbot returns a standardized fallback response indicating that the requested information cannot be confirmed from the available documents. In this way, the prompt functions not only as a generation instruction, but also as a governance mechanism supporting accuracy, transparency, and human oversight.

3.5.3. Model runtime and deployment

The HR chatbot relies on locally hosted large language models executed through an on-premise runtime. All inference is performed within institutional infrastructure, without external network access. The system supports CPU-based execution and optional GPU acceleration depending on available hardware.

To maintain predictable operational behaviour, generation parameters such as output length, temperature, and time limits are explicitly bounded. This allows pilot deployments to balance answer quality with acceptable latency under offline inference conditions. Figure 2 illustrates the HR chatbot interface.

Figure 2: User interface of the offline HR chatbot PoC with configurable LLM settings and response-time telemetry



3.6. Governance and Safety Controls

Because the system is intended for a regulated institutional environment, governance and safety controls are implemented at the application level rather than treated as optional add-ons. These controls are designed to ensure that outputs remain verifiable, bounded, and compatible with institutional expectations of accountability and human oversight.

A central design principle is traceability. Retrieved passages are displayed alongside system outputs so that users can inspect the evidentiary basis of both search results and generated answers. This improves interpretability and reduces reliance on opaque model behaviour.

Additional safeguards are implemented through output constraints, including language restriction, structured answer formatting, and fallback behaviour when supporting evidence is insufficient. Together, these measures reduce the scope for hallucinated or speculative responses and align the assistants with a document-grounded support role rather than an autonomous advisory role.

3.6.1. Data protection rules

The HR chatbot is limited to providing general, policy-based information. It is explicitly prohibited from generating personalized responses about identifiable individuals, including employment status, salary, disciplinary measures, or individual benefits. Queries requiring personal data are rejected or redirected to official HR channels.

3.6.2. Logging and auditability

To support accountability, the system maintains minimal operational logs for each interaction. Logged metadata includes a timestamp, anonymized user identifier, a hash of the query, and references to the retrieved sources. Full conversation content and personal data are not retained, thereby supporting auditability while minimizing data retention risk.

3.6.3. Access control

Access to the chatbot and its document indices is controlled through role-based permissions. Users are granted access only to document collections relevant to their organizational role or unit. This separation reduces the risk of unintended exposure of sensitive internal materials and supports the principle of least privilege.

3.7. Deployment and Operational Considerations

The proposed architecture is intended for deployment on standard server or workstation hardware within a controlled institutional network. All processing steps—document parsing, embedding generation, indexing, retrieval, and generation—are executed locally and without Internet connectivity.

Operationally, the system supports offline execution, incremental indexing, configurable model selection, and separation between data, models, and application logic. These characteristics make it suitable for pilot deployment in central banking environments, where experimentation with AI systems must be carefully balanced against confidentiality, security, and governance requirements.

At the time of writing, the system was transitioning from proof-of-concept to pilot use within a controlled institutional environment. Although not yet a production-grade platform, it already supports real document collections and practical internal use cases, thereby providing an empirical basis for further refinement and evaluation.

4. Evaluation and discussion

4.1. Dataset and Document Corpus Description

The proof-of-concept system was evaluated on an internal document corpus from the Central Bank of Bosnia and Herzegovina. The broader document repository contains approximately 4,000 files, while the evaluated subset used for indexing experiments consisted of 1,318 documents. The corpus primarily includes documents in PDF and Microsoft Word formats, such as press releases, laws and bylaws, regulatory documents, internal reports and procedures. Document creation dates span a period from approximately 2010 to 2025.

Figure 3: Text segmentation settings for semantic indexing: chunk size and overlap

For semantic search purposes, documents were divided into overlapping text chunks of approximately 900 characters, using an overlap of 150 characters to preserve contextual continuity between segments as shown in Figure 3.

This resulted in approximately 7247 indexed text segments, from 1,318 files with an average of 5.5 chunks per document.

4.2. Evaluation Approach and Scope

The evaluation combines three complementary components: (i) system-level performance measurements (indexing time and query latency), (ii) a structured query–relevance assessment of retrieval modes, and (iii) a small-scale qualita-

tive evaluation of answer grounding for the HR chatbot. Because the system is at pilot stage and operates in a security-constrained institutional environment, the evaluation is intended to provide structured feasibility evidence rather than benchmark-level claims. The assessed use cases were designed to reflect realistic information needs within a central bank, including regulatory references, internal procedures, supervisory topics, and analytical reports stored across heterogeneous document repositories. The evaluation therefore prioritises operational usefulness, interpretability, and governance compatibility, while still incorporating limited systematic retrieval and answer-quality assessment.

4.3. Indexing and Performance Characteristics

The initial indexing and embedding generation process was executed entirely on-premise using CPU-based computation. For the evaluated corpus (approximately 1,300 documents resulting in 7,247 indexed text chunks), the full indexing process required approximately two hours on standard institutional hardware (Intel® Xeon® CPU E5-2620 v4, 12 vCPUs, 16 GB RAM).

Indexing time was primarily influenced by the selected embedding model and chunking configuration. Lightweight multilingual models demonstrated faster embedding generation and lower memory consumption, while higher-capacity models required longer processing times but provided improved semantic representation quality.

To optimize operational usability, embeddings were cached locally in a FAISS index, allowing subsequent re-indexing to be performed incrementally when document changes occurred. This significantly reduced maintenance overhead and enabled practical updates without full recomputation.

The indexing process represents a one-time or infrequent computational cost, amortized over query-time operations, which remained suitable for interactive use as discussed below.

4.4. Semantic Document Search Characteristics

At query time, average response latency remained below 0.5 seconds for exact phrase search and below 2 seconds for semantic and hybrid search modes, supporting interactive use in day-to-day operational contexts. The evaluation is based on observations from a pilot-stage deployment of the system, reflecting its transition beyond a standalone proof-of-concept.

4.5. Comparative Behaviour of Search Modes for Semantic Document Search

The evaluation revealed clear functional distinctions between the three supported search modes, each addressing different categories of information needs.

Exact phrase search performed reliably in cases where users knew the precise terminology or phrasing used in documents such as references to formal titles, legal provisions or specific regulatory expressions. Its deterministic behaviour and high precision make it well suited for compliance-oriented tasks, but its reliance on exact wording limits recall when terminology varies across documents or organizational units.

Semantic similarity search demonstrated significant advantages in exploratory and analytical scenarios. Queries expressed in descriptive or conceptual terms frequently retrieved relevant document fragments even when no direct lexical overlap was present. This behaviour was particularly valuable for retrieving internal analyses, reports, and background materials authored using heterogeneous vocabulary. However, purely semantic ranking occasionally surfaced conceptually related but contextually less relevant results, especially for short or highly generic queries.

The hybrid search mode proved most effective for day-to-day professional use. By combining semantic similarity with lexical signals, the hybrid approach balanced recall and precision, promoting results that were both conceptually relevant and lexically grounded. This re-ranking strategy reduced the incidence of semantically plausible but practically irrelevant results and aligned well with the expectations of expert users accustomed to traditional search systems.

4.6. Performance and Resource Considerations

While embedding generation is computationally intensive during initial indexing, this cost is incurred offline and amortized over subsequent search operations through the use of persistent local caches. The model-specific caching mechanism significantly reduced re-indexing time when documents remained unchanged, enabling practical maintenance workflows even for moderately large document collections. Query-time performance was observed to be responsive for interactive use, as similarity computations are limited to precomputed embeddings and do not require external model calls. These observations suggest that transformer-based semantic search can be practically deployed in offline, on-premise environments without specialized hardware, provided that indexing

operations are managed appropriately. The categories in Table 1 reflect empirical observations during pilot indexing and querying under the described hardware setup, complemented by model documentation.

Table 1: Characteristics and deployment trade-offs of evaluated multilingual embedding models in the semantic search pipeline

Model name	Approx. memory usage	Relative speed	Semantic quality	Typical use case
paraphrase-multilingual-MiniLM-L12-v2	Low	High	Moderate	Large document collections, frequent indexing, limited hardware
paraphrase-multilingual-mpnet-base-v2	Medium	Medium	High	Analytical queries requiring better semantic precision
BAAI/bge-m3	High	Low	Very high	Complex semantic queries, smaller collections, powerful hardware

4.7. Structured Evaluation of Retrieval Modes

To complement the qualitative observations described above, a small structured evaluation of the retrieval modes was conducted using a set of 30 representative user queries reflecting typical institutional information needs, including regulatory references, internal procedures, supervisory topics, and analytical reports.

For each query, the top three results returned by the three retrieval modes (exact phrase search, semantic similarity search, and hybrid search) were manually inspected. Relevance was assessed against the query intent and the returned document fragment. A result was labelled relevant if it contained information directly answering the query or clearly supporting the requested topic. The relevance judgments were performed manually by the authors based on inspection of the retrieved passages and the underlying documents. To improve evaluation consistency, a subset of ten queries was independently reviewed by a second evaluator familiar with the document corpus. The resulting relevance assessments were broadly consistent with the initial labels. Minor differences in interpretation were resolved through discussion. Because the study is intended as a pilot-stage feasibility assessment, relevance judgments were binary and limited to the top three results per query. Although the query set is relatively small, it was designed to reflect representative institutional information needs observed during pilot system use.

Table 2 reports the resulting Success@3 values, defined as the share of queries for which at least one of the top three returned fragments was relevant, a metric com-

monly used in information retrieval evaluation to assess retrieval success within the top-k results (Manning et al., 2008). In addition to Success@3, Precision@3 was computed to provide a complementary perspective on ranking quality by measuring the proportion of relevant fragments among the top three returned results. Precision@3 was derived from the manually assigned binary relevance labels for the top three retrieved fragments across the evaluated query set.

Table 2: Retrieval performance across search modes measured by Success@3 and Precision@3

Retrieval mode	Queries with ≥1 relevant result in top 3	Success@3	Precision@3
Exact phrase search	17/30	56.7%	42.0%
Semantic similarity search	22/30	73.3%	56.0%
Hybrid search	26/30	86.7%	68.0%

While Success@3 captures whether at least one relevant result appears among the top-ranked fragments, Precision@3 reflects the density of relevant information presented to the user within the top results.

The results indicate that the hybrid retrieval mode provided the most reliable performance across diverse query types. These findings are consistent with the qualitative observations presented earlier. Exact phrase search performs well when the user knows the precise wording used in the documents, while semantic similarity search improves recall when terminology varies. The hybrid re-ranking approach combines these advantages by preserving lexical precision while promoting semantically relevant passages, resulting in more stable retrieval performance in professional document collections.

4.8. HR Chatbot Runtime Characteristics

In addition to semantic document search, the HR chatbot proof-of-concept was evaluated with respect to response latency under offline, CPU-based inference conditions. The chatbot reuses the same FAISS-based vector index and embedding store as the document search assistant; however, in the chatbot pipeline, the retrieval stage includes not only vector lookup but also query embedding computation and preparation of retrieved passages for prompt construction. Under a representative configuration (Model: qwen2.5:1.5b, TOP_K = 3, temperature = 0.2), the total response time was 53.5 s, composed of 6.5 s for retrieval-stage processing and 47.0 s for local LLM generation. Overall, end-to-end latency was dominated by LLM inference, while retrieval contributed a smaller but non-neg-

ligible share. These measurements highlight a practical quality–latency trade-off and suggest that pilot deployments should tune model size, TOP_K, and generation limits to balance answer quality with acceptable response times.

4.9. Mini quality evaluation for HR Chatbot

To complement latency measurements, a small-scale qualitative evaluation of answer grounding was conducted on a controlled set of 30 HR questions covering: (i) directly answerable policy queries, (ii) procedural "how-to" queries, (iii) exception cases, and (iv) intentionally out-of-scope questions designed to trigger fallback behaviour. For each query, the retrieved TOP_K passages and the generated response were logged and manually assessed against the retrieved evidence and the intended answer scope. The qualitative labels were assigned manually by the authors based on inspection of the retrieved evidence and the generated answer. Responses were labelled as fully supported, partially supported, or unsupported. In addition, cases where fallback behaviour was expected were reviewed to assess whether the standardized fallback response was appropriately triggered. Fallback behaviour was considered appropriate when the retrieved passages did not provide sufficient evidence to support a reliable policy-based answer within the scope of the indexed HR documents. In such cases, the chatbot was expected to return the predefined fallback response rather than generate a speculative answer. Table 3 summarises the results.

Table 3: Summary of HR chatbot answer-quality and fallback evaluation

Evaluation metric	Result
Fully supported responses	22/30 (73.3%)
Partially supported responses	6/30 (20.0%)
Unsupported responses	2/30 (6.7%)
Correct fallback triggers	4/10 (40.0%)

Overall, the results suggest that the implemented governance constraints reduce the incidence of unsupported answers under offline inference conditions. This corresponds to a fallback precision of 40.0%. At the same time, the comparatively low fallback precision indicates that fallback triggering remains insufficiently calibrated in the present proof-of-concept and is sensitive to retrieval quality, document coverage, and prompt behaviour. This limitation should therefore be interpreted not as a failure of the governance concept itself, but as evidence that retrieval thresholds and fallback policies require further refinement before production deployment.

Beyond the qualitative grounding assessment, informal pilot feedback from HR staff who tested the chatbot indicated that it was perceived as a useful solution for routine policy-related questions. As this feedback was not collected through a structured user-study design, it is reported here only as indicative practitioner feedback.

4.10. Impact on Access to Institutional Knowledge

From an operational perspective, the PoC system demonstrated clear potential to enhance access to institutional knowledge embedded in unstructured documents. Users were able to retrieve relevant information without prior knowledge of document structure, storage location, or exact phrasing. The chunk-based retrieval strategy further enabled direct access to specific passages within documents, reducing the need to manually scan entire files.

This capability is particularly relevant in central banking environments, where institutional memory is often distributed across long-lived document collections and staff turnover can lead to loss of contextual knowledge. By enabling meaning-based retrieval across heterogeneous sources, the system supports more efficient knowledge reuse and reduces dependency on informal expertise or personal familiarity with document repositories.

4.11. Security and Governance Implications

Data protection concerns and third-party dependency have been identified across different banking AI applications, reinforcing the relevance of fully controlled, offline-capable architectures (Sadok et al., 2022; Villar & Khan, 2021). A central outcome of the evaluation is the confirmation that advanced NLP-based document search can be implemented without compromising security or data sovereignty. The absence of Internet connectivity during normal operation eliminates exposure to external services, reduces attack surface, and simplifies compliance with data protection and confidentiality requirements.

From a governance perspective, the system's transparent retrieval logic—particularly the hybrid re-ranking approach—supports interpretability and user trust. Retrieved text fragments are explicitly displayed, allowing human users to validate relevance rather than relying on opaque automated decisions. This design choice aligns with central banking principles emphasizing accountability, explainability, and human oversight in the use of AI technologies.

4.12. Limitations and Practical Considerations

While the PoC demonstrates clear benefits, several limitations were identified during evaluation. The current implementation relies on periodic re-indexing to incorporate document changes, which may limit suitability for environments with highly dynamic document repositories. In addition, access control is enforced at the application level rather than at a fine-grained document or fragment level.

The quality of semantic retrieval is influenced by the choice of embedding model, which involves trade-offs between computational cost and representational richness. Although multilingual models perform robustly across diverse document types, domain-specific fine-tuning was not explored in this PoC and represents an area for potential future enhancement.

Although the proposed architecture is likely transferable to other central banking environments, its practical deployment will depend on local factors such as corpus size, document heterogeneity, language distribution, access-control requirements, and available hardware resources. Larger institutions may require more scalable indexing strategies, tighter integration with document management systems, and more granular authorization models. Therefore, the present findings should be interpreted as evidence of feasibility in a medium-scale, multilingual, security-constrained environment rather than as directly generalizable performance benchmarks for all central banks.

The manual evaluation design improves interpretability in a pilot setting, but it remains sensitive to evaluator judgment and should therefore be expanded in future work through broader expert-based assessment.

4.13. Implications for Central Banks and Transferability

Although the system was developed within the Central Bank of Bosnia and Herzegovina, several design principles may be transferable to other central banks and regulated public-sector institutions operating under similar constraints. First, offline-first deployment is appropriate where internal document repositories contain sensitive or confidential information that cannot be processed through external services. Second, multilingual embedding models are particularly useful in document environments where language use is heterogeneous and cannot be reduced to English-only workflows. Third, hybrid retrieval offers a pragmatic balance between lexical precision and semantic recall, which is especially rel-

evant in professional institutions where both formal terminology and conceptual search matter. Finally, transparent display of retrieved passages, bounded generation, and explicit fallback behaviour support interpretability and human oversight, making such systems more compatible with internal governance expectations. At the same time, transferability depends on local factors such as corpus size, hardware capacity, language distribution, access-control requirements, and integration with document management systems.

4.14. Discussion Summary

Overall, the evaluation indicates that the proposed PoC system effectively addresses a well-defined operational gap in central banking environments: the need for secure, offline, and semantically meaningful access to unstructured internal documents. By integrating classical information retrieval techniques with modern sentence embeddings in a transparent and controllable architecture, the system demonstrates that AI-based document search can be responsibly applied within highly regulated institutional contexts.

The findings support the feasibility of extending the PoC into a production-grade family of offline assistants, provided that future work addresses scalability, access control, retrieval quality, and integration with existing document management and internal support infrastructures.

5. Conclusion

This paper presented two proof-of-concept NLP assistants designed to improve secure internal knowledge access in the Central Bank of Bosnia and Herzegovina: a semantic document search system and a document-grounded HR chatbot built on a shared offline architecture. The paper does not claim methodological novelty in retrieval or language generation. Instead, its contribution lies in showing how established NLP techniques can be adapted to the governance, confidentiality, and operational requirements of a central bank through an offline on-premise design.

The evaluation results indicate that hybrid lexical–semantic retrieval provided the most reliable performance across the assessed internal query set, while the HR chatbot demonstrated the feasibility of source-grounded conversational support under bounded offline inference conditions. At the same time, the findings should be interpreted as pilot-stage feasibility evidence rather than as benchmark-

level performance claims. In particular, retrieval evaluation remains limited in scale, and fallback behaviour in the HR chatbot requires further calibration.

More broadly, the study suggests that offline NLP assistants can provide a practical and governance-compatible approach to institutional knowledge access in central banking environments. The combination of offline semantic retrieval, transparent source display, bounded generation, and fallback behaviour may serve as a useful deployment pattern for other regulated public-sector institutions facing similar confidentiality and operational constraints.

Future work should extend the evaluation to larger and more diverse document repositories, strengthen retrieval assessment with broader expert-based relevance judgments, improve fallback calibration in the HR chatbot, and address production-oriented requirements such as finer-grained access control, scalability, and integration with existing document management systems.

References

1. Araujo, D., Doerr, S., Gambacorta, L., & Tissot, B. (2024). Artificial intelligence in central banking (*BIS Bulletin* No. 84). *Bank for International Settlements*. https://www.bis.org/publ/bisbull84.pdf
2. Bernik, N., & Šprajc, P. (2025). Use of chatbots in human resource management for more efficient knowledge sharing: A systematic literature review. *Organizacija*, 58(4), 283–296. https://doi.org/10.2478/orga-2025-0024
3. Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Cooke, F. L., Decker, S., Denisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A. K., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., Rogelberg, S., Saunders, M. P., Tung, R. L., & Paluch, R. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, 33(3), 606–659. https://doi.org/10.1111/1748-8583.12524
4. Cimpeanu, I. A., Dragomir, D. A., & Zota, R. D. (2023). Banking chatbots: How artificial intelligence helps the banks. *Proceedings of the International Conference on Business Excellence*, 17(1), 1716–1727. https://doi.org/10.2478/picbe-2023-0150
5. Doerr, S., Gambacorta, L., & Garralda, J. M. S. (2021). Big data and machine learning in central banking (BIS Working Papers No. 930).
6. Gambacorta, L., Kwon, B., Park, T., Patelli, P., & Zhu, S. (2024). CB-LMs: Language models for central banking. BIS Working Papers (forthcoming).
7. Grigorescu, A. E. (2024). Artificial intelligence in central banking. In *Proceedings of the 18th International Conference on Business Excellence 2024*.
8. Koti, K. (2024). The role of artificial intelligence in shaping customer experiences in the banking sector. Library of Progress – Library Science, *Information Technology & Computer*, 44(3).
9. Li, B., & Cheng, Y. (2025). ChatGPT in human resource management: A systematic review of influential factors, processes, and outcomes. *Heliyon*, 11(10), e44048. https://doi.org/10.1016/j.heliyon.2025.e44048
10. Majumder, S., & Mondal, A. (2021). Are chatbots really useful for human resource management? *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-021-09834-y
11. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
12. Sadok, H., Sakka, F., & El Maknouzi, M. E. H. (2022). Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, 10(1), 2023262. https://doi.org/10.1080/23322039.2022.2023262

13. Soundenkar, S., Bhosale, K., Jakhete, M. D., Kadam, K., Chowdary, V. G. R., & Durga, H. K. (2024). AI-powered risk management: Addressing cybersecurity threats in financial systems. *Library of Progress – Library Science, Information Technology & Computer*, 44(3).
14. Vučinić, M., & Luburić, R. (2024). Artificial intelligence, fintech and challenges to central banks. *Journal of Central Banking Theory and Practice*, 13(3), 5–42.
15. Villar, A. S., & Khan, N. (2021). Robotic process automation in banking industry: A case study on Deutsche Bank. *Journal of Banking and Financial Technology*, 5(1), 71–86. https://doi.org/10.1007/s42786-021-00035-4 /.