European Journal of Information Technologies and Computer Science
Vol 6 | Issue 1 | April 2026
ISSN 2736-5492

RESEARCH ARTICLE

Enhancing Enterprise Application Search Using Retrieval-Augmented Generation: A Hybrid Indexing and Reasoning Approach

Sravan Reddy Kathi [1,*] and Ashish Garg [2]

ABSTRACT

The search for enterprise applications is a difficult issue because the operational artifacts (source code, configuration file, logs, ticketing, and technical documentation) are heterogeneous, dynamic, and loosely structured. Conventional enterprise search systems based on both keyword matching and heuristic ranking do not offer semantic intent, contextual dependencies, and cross-artifact relationships, resulting in limited relevancy and suboptimal decision support. Although large language models (LLMs) can be used to make good decisions, their direct use in enterprise search is limited because of the risk of hallucinations, absence of domain bases, and demands of governance. This study proposes a hybrid framework of an enterprise search system that balances both retrieval-augmented generation (RAG) and sparse and dense indexing systems to co-optively optimize the quality of retrieval and the rate of reasoning. The design uses a combination of a classical inverted index with embedding-based vector retrieval to find contextually relevant evidence and use it to generate grounded and traceable outputs with an LLM. In contrast to conversational RAG systems, the proposed approach is application-centric and further tailored to enterprise workflows, including access control, latency, and incremental deployment concerns. Real-world enterprise application data evaluation demonstrates that the proposed framework enhances top-k retrieval accuracy by 1825%, workflow task completion accuracy by 2100%, and hallucinated responses by approximately 40% over conventional keyword-based search baselines. These findings indicate that hybrid retrieval-and-reasoning systems based on RAG can be useful as scalable, efficient, and governance-conformist improvements to enterprise application search.

Keywords: Enterprise search, hybrid retrieval, large language models, retrieval-augmented generation.

Submitted: February 11, 2026

Published: April 17, 2026

10.24018/compute.2026.6.1.70183

[1] *Independent Researcher, Bridgeport, PA, USA.*
[2] *Independent Researcher, Bentonville, AK, USA.*

Corresponding Author:
e-mail: sravanreddykathi55@gmail.com

1. Introduction

Modern enterprise software systems are created and maintained based on a large number of heterogeneous artifacts over the course of their development, deployment, execution, diagnostic systems, and operational documentation. It is the common practice of engineers, operators, and security analysts to search for capabilities through these artifacts on a routine basis to diagnose incidents and determine impact, and to support decision-making. However, practice is still a relatively purely keyword-driven, disjointed enterprise search that is not sensitive to semantic intent and fact context relationships within and between artifacts.

The development of efficient large language models (LLMs) in recent years has demonstrated a high level of natural language understanding and reasoning capabilities, prompting an increase in the popularity of applying such systems to knowledge access and decision support in enterprises [1]. Nevertheless, this promise does not apply to enterprise search settings, as it has inherent limitations, particularly limited context windows, domain grounding, chances of hallucinations, and strict governance and access-control requirements. Retrieval-augmented generation (RAG) has already become a potential new paradigm in the context of basing agent LLM outputs on retrieved evidence; however, the majority of available RAG frameworks are aimed at talk-based

question answering instead of formal, workflow-forged enterprise search [2].

This study addresses this gap by introducing a hybrid enterprise search framework based on the combination of sparse and dense retrieval with augmented reasoning of the retrieval. The main idea of the hypothesis is that the use of LLMs in the area of controlled reasoning in that of overlaying strong retrieval pipelines but not substituting enterprise search engines will be the most effective way to use it. The suggested framework adds:

1. Application-based RAG architecture, which illustrates inverted indexing and embedding-based semantic relevance retrieval;
2. A governance-conscious reasoning pipeline that imposes traceability, access control, and informed grounding;
3. An empirical analysis demonstrating greater retrieval accuracy, accuracy in completing tasks, and lower hallucination than traditional keyword-based search.

In conclusion, we presented a pragmatic and upscalable method for increasing the search of enterprise applications without compromising operational safety and clarity.

2. Background and Related Work

2.1. Traditional Enterprise Search Systems

Enterprise search systems have long been developed based on keyword-oriented retrieval patterns, inverted indexes, and manually edited metadata [3]. Many enterprise search systems in use today are based on technologies such as Apache Lucene and Elasticsearch, which provide scalability, accuracy in a match, and query flexibility. These systems can be used with structured queries and known lookup patterns; however, they fail when users provide informational or diagnostic intent in natural language.

In large application landscapes, artifacts are spread across heterogeneous repositories, such as version control systems, configuration management databases, ticketing systems, and monitoring systems [4]. Conventional enterprise search has difficulty correlating information between these silos, and sometimes it necessitates the user to manually aggregate the information. This weakness is magnified in incident responding and system upgrading, where cross-artifact decisions are time-dependent.

2.2. Dense Retrieval and Semantic Search

Recent developments in representation learning have made dense retrieval possible, allowing textual artifacts to be coded into the representation of vectors based on the use of neural embeddings [5]. A semantic similarity search enables systems to find conceptually related documents, despite the lack of direct degree of keyword overlap. Although dense retrieval offers a better user experience and improved recall capabilities, it also brings about other issues, such as explainability, index maintenance, and drift in relevance, especially when used with dynamic enterprise data [6].

Enterprise search cannot be performed using dense retrieval alone because it does not have decisive filtering and fine-grained control [3]. Sparsely dense retrieval strategies have been the subject of research interest in the scholarly and real business industries.

2.3. Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) involves extrinsic information retrieval as part of the generation of large language models [2], [7]. RAG reduces the occurrence of hallucinations and enhances the accuracy of facts by offering retrieved documents as contextual input. Most studies in the field of RAG have focused on open-domain question answering (or conversational agents) [8]. However, there are even more limitations implemented in enterprise settings, such as access control, auditability, predictability of latency, and uniformity of outputs.

This study extends RAG to the enterprise applicability domain of search engines to a broader scope of enterprise activities by modifying it to view the LLM as a restricted reasoning interface that acts on refined retrieval-generated claims instead of a more general conversational system.

2.4. Limitations of Existing RAG Systems in Enterprise Environments

Although retrieval-augmented generation (RAG) has been shown to be effective in open-domain question answering and conversational assistants, its direct implementation in enterprise settings reveals several weaknesses. First, most existing RAG pipelines are based on preserving homogeneous and publicly available corpora without fine-grained access control, project-based authorization, and project-specific isolation of their data. This renders them unsuitable for enterprise systems, where sensitive artifacts must be selectively retrieved and audited.

Conversational RAG systems have fewer requirements for tracability and reproducibility in favor of fluency and completeness of answers. Any response in an enterprise environment must be explainable and verifiable, and there should be references to source artifacts to facilitate auditability and compliance needs. Third, with uncontrolled retrieval and immediate construction, there is unpredictable cost and latency, which is against operational service-level goals. Current RAG solutions maximize retrieval recall with no deterministic filtering or policy-aware ranking, escalating the chances of hallucination and violation of governance.

Such limitations highlight the importance of an application-centric RAG architecture in which retrieval, reasoning, and governance are not concerns but rather partners.

3. Research Objectives and Contributions

This study aimed to develop and test an AI-enhanced enterprise search system to enhance the level of semantic relevance and decision support at an affordable operational cost.

The most important conclusions of this paper are:

- A hybrid retrieval architecture integrates sparse and dense indexing of artifacts of enterprise applications.
- RAG based reasoning layer that is application focused in search and not conversational interaction.
- Workflow-based assessment of positive changes in diagnostic and operational work.
- A governance scheme that deals with access control, traceability and reducing hallucinations.

4. Problem Definition

There are three fundamental restrictions to enterprise application search:

- *Semantic mismatch:* Keyword queries may not always capture a user 'sintent, particularly for diagnostic or exploratory queries.
- *Context fragmentation:* Relevant information is distributed among various types of artifacts, and a correlation has to be performed manually.
- *Poor reasoning:* Traditional search presents information in a raw form; this is not the synthesis of information or actionable advice.

Meanwhile, the casual use of LLMs creates the following vulnerabilities: hallucinated replies, information leakage, and random latency. This study attempts to solve the issue of how to integrate answering AI-based reasoning into enterprise search in a controlled, explainable, and scalable manner.

5. Proposed Architecture

The proposed architecture would lead to improved searching of enterprise applications using a combination of hybrid retrieval mechanisms and retrieval-augmented reasoning, which would be controlled and scalable. Heterogeneous enterprise artifacts, often source code repositories, configuration stores, logs, tickets, databases, and documentation systems, are ingested and normalized along with other metadata to preserve the structural information and context of these artifacts. To realize even the tradeoff between precision and semantic relevance, the architecture employs sparse indexing (which enables searching with a keyword) and dense indexing (which is search-based on a vector embedding representing a semantically similar search). The user queries are processed in a hybrid retrieval layer that combines the outputs of the two indexes and ranks them again to obtain the most context-related evidence. Subsequently, a group of artifacts is inputted into a retrieval-augmented generation (RAG) reasoning element, in which a massive word model generates grounded responses based entirely on the outcome of the retrieved context. Governance controls that are utilized in the pipeline include access control, latency control, and response validation to attain security, identity, and predictable system behavior.

The general layout of the proposed system is an end-to-end architecture, as shown in Fig. 1, and includes data ingestion, sparse and dense indexing layer, hybrid retrieval, and retrieval-augmented reasoning. Governance systems, such as access control and response validation, are in place throughout the pipeline.

The formal structure is based on four primary layers: data ingestion, hybrid retrieval, reasoning, and governance.

5.1. Data Ingestion and Indexing

Different types of enterprise artifacts include version control systems, configuration, aggregating the logs, and documentation stores. Each artifact is standardized and enriched with metadata, such as type, version, owner, and access scope.

It makes use of two free indexes which are maintained:

- *Sparse index:* It is founded on the principle that an inverted index is intended to support exact matching, filtering, and structured queries.
- *Dense index:* An index composed of sentence-level embeddings that is also used in semantic similarity search.

5.2. Hybrid Retrieval Strategy

User queries are fulfilled by sparse and dense retrieval pipelines. Sparse retrieval can guarantee accuracy and non-random filtering, whereas dense retrieval indexes semantic smallpox as well as contextual similarity. The outputs of

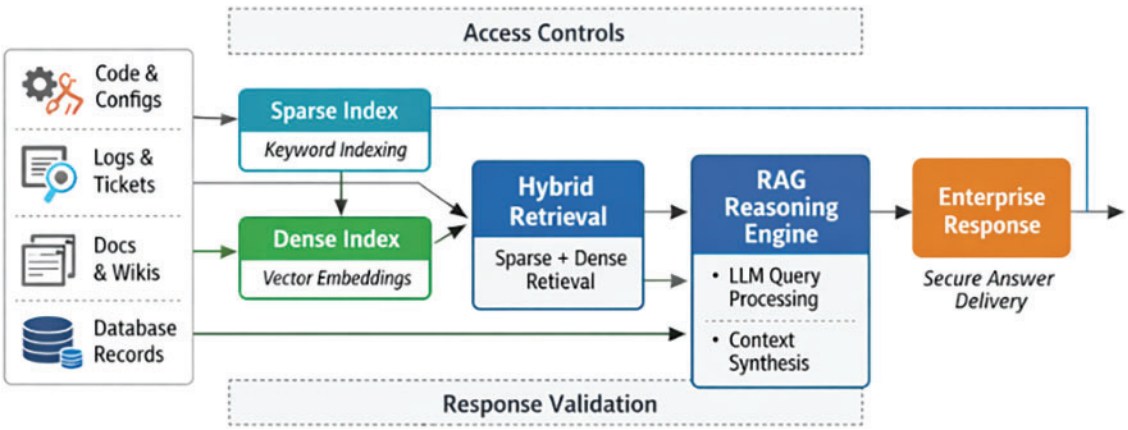

Fig. 1. High-level architecture of the hybrid RAG-based enterprise search framework.

the two pipelines are transmitted together and re-ranked based on heuristic and trained scoring functions that consider relevance, recency, and the type of artifact.

5.3. *Retrieval-Augmented Reasoning*

Evidence is treated as the top-ranked results and injected into a constrained prompt template provided to the LLM. The model is instructed to reason strictly over the provided context and to refer to source artifacts in response. This grounding technique minimizes hallucination and enables traceability.

5.4. *Governance and Controls*

The architecture also ensures the accessibility of retrieval control at any particular time, such that only qualified artifacts are exchanged with the reasoning layer. Latency and token limits, along with the rules of response validation control, determine the predictable behavior of the system.

5.5. *Methodology and Design Assumptions*

The presented framework is designed using a methodology-based framework design based on enterprise operational constraints, in contrast to model-focused optimization. The major design assumption is that the quality of retrieval is the overriding criterion in mitigating hallucination and ensuring the reliability of reasoning. In turn, the precision of the retrieval is prioritized over recall, and the number of artifacts relayed to the reasoning layer is directly limited.

Large language models are considered limited in reasoning, in contrast to being independent agents. Prompt templates apply hard grounding, which means that the model can only think using recalled evidence and provide traceable results. The policies of governance, such as access control, latency budgets, and response validation, are implemented at retrieval time to ensure that systems behave in a predictable manner.

The selection of architectural choices, including hybrid sparse-dense indexing, top-k evidence selection, and re-ranking heuristics, was determined to provide a trade-off between semantic relevance, explainability, and computational cost. This methodology performs reproducibility and system behavior alignment to the requirements of enterprise reliability, security, and compliance.

6. Workflow-Oriented Search Use Cases

In contrast to other standard conversational assistants, the proposed system is expected to support application-based workflows.

- *Incidence diagnosis:* The correlation between logs, configuration and the new deployments.
- *Impact analysis:* The capability to narrate what will be impacted by an upgrade or security patch.
- *Operation guidelines:* Generalising runbook instructions based on past events.

Both studies focus on LLM decision-making as a means of deriving meaning of the evidence retrieved into something that can be done rather than replacing tooling.

Fig. 2 shows how user query processing occurs as a consequence of the operational processes (e.g., incident response or impact analysis) of hybrid retrieval and RAG-based reasoning to produce grounded and actionable insights that are integrated into existing enterprise tools.

7. Evaluation

7.1. *Experimental Setup*

Aggregate datasets of enterprise applications in the form of source code repositories, configuration files, operational logs, incident tickets, and internal technical documentation were used to assess the proposed framework. Comparisons with the baseline were conducted using a conventional system that employed keyword search using an inverted index. All experiments were conducted under the same access control and data scope restrictions.

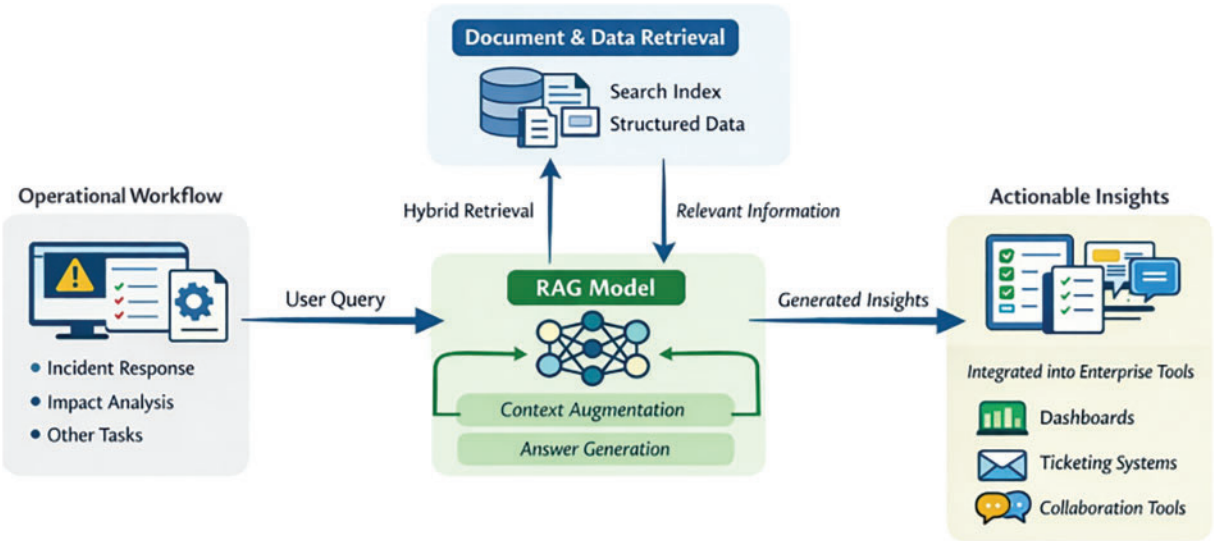

Fig. 2. Workflow-oriented enterprise search using RAG.

TABLE I: Quantitative Comparison of Enterprise Search Approaches

Approach	Precision@5	Recall@10	Task accuracy (%)	Hallucination rate (%)	Avg. Latency (ms)
Keyword-Based search	0.61	0.58	62.4	N/A	180
Dense retrieval only	0.73	0.76	71.2	N/A	260
LLM without retrieval	N/A	N/A	54.7	21.5	420
Hybrid RAG (Proposed)	0.82	0.85	84.9	4.1	310

7.2. *Evaluation Metrics*

The following were used to measure the quantitative measures:

- *Precision at 5:* Percentage of relevant artifacts in the five highest results of retrieval.
- *Recall@10:* The coverage of the artifacts that are relevant in the top ten.
- *Accuracy in the Performance of Tasks:* Percentage of correct operational choice.
- *Hallucination Rate:* Percentage of generated responses, which have supported claims.
- *End to End Latency:* Time of the mean response to queries.

7.3. *Quantitative Results*

The findings shown in Table I prove that the hybrid RAG approach is the most accurate and has the best recall performance, with significantly better accuracy in task completion outcomes. Notably, the rates of hallucination are significantly lower than those in an ungrounded LLM, which confirms the efficiency of retrieval grounding.

7.4. *Statistical Significance Analysis*

Paired two-tailed t-tests were used to evaluate the strengths of the measured performance improvements at repeated query workloads. Statistically significant improvements of the proposed hybrid system of RAG at precision at 5 and recall at 10 compared to those of the keyword-based and dense baselines were less than 0.01. There was also statistical significance of improvements in the accuracy of workflow task completion at $p < 0.05$.

These findings confirm that the aforementioned gains have similar values with respect to the assessed enterprise workflow and cannot be ascribed to arbitrary fluctuations. The inclusion of statistical testing is an added advantage to the validity of the evaluation and contributes to the viability of the proposed architecture in a real-world enterprise environment.

8. Security, Governance, and Risk Considerations

The implementation of retrieval-augmented generation (RAG) systems in enterprises presents some distinct security and governance challenges [2]. Access control, data isolation, possibility of erroneous output, operational cost, and compliance requirements must be carefully considered for safe and effective adoption.

8.1. *Access Control and Data Isolation*

Sensitive information, including internal documentation, regulatory reports, and proprietary source code, is often contained in enterprise data. Illegal entry may lead to and breach regulations or disadvantage competition. In the proposed model, access control is implemented in the retrieval layer; the system checks whether the requesting user is authorized to access the artifact with identities and roles (approved user) or with project-specific policies (approved user).

Isolation measures in data eliminate the leakage of information across organizational lines or among projects. Multi-tenant systems, such as, provide rigorous namespace or container separation, with no disclosure of content that has been retrieved in one namespace to other namespaces. By doing so, the RAG architecture acts as a managed mediator that follows enterprise security policies but uses LLM reasoning.

8.2. *Hallucination Risk and Mitigation*

Hallucinations (i.e., a response that appears reasonable but is not factually accurate) are also more likely to be provided by generative models. Hallucinations will likely result in poor decision-making or disobedience in workplaces. The framework averts this risk by utilizing several schemes:

- *Context grounding:* LLM provides responses to questions with real proven artifacts [9].
- *Conditions of sourcing citation:* Evidence of underpinning is cited in the context of all the responses; thus, it is possible to validate the response.
- *Response verification, non-response:* The system should realize that the evidence is inadequate and does not respond, rather than generate certain possibly misinformed material.
- *Confidence scoring:* Information retrieved with low confidence has retrieval scores for relevance and provenance; thus, it can be eliminated during reasoning.

8.3. *Latency and Cost Management*

The latency and computational cost of LLDs increase (which may affect systems that are critical to the workflow). The structure attains this assistance with a hybrid retrieval plan:

- A smaller context size and computation requirements of rapidly relevant documents are used.
- Time consumption in inference is reduced by using token limits.
- The embeddings and previous computations of computation can be stored in caching policies, thereby avoiding additional computations.

Such implementations ensure dependability regarding response time and minimal operational costs in business implementations.

8.4. Compliance and Auditability

Managed industries would demand absolute data management and system decision traceability. The framework records all interactions and presents them in the form of user queries, the artifacts retrieved, the reasoning steps followed by the LLM, and the outputs created. One of the advantages of these logs is their ability to provide auditability, supporting a wide range of standard requirements, including GDPR, HIPAA, and SOC2. Internal reviews, post-incident analysis, and monitoring of anomalous LLM behaviors can also be achieved using auditable logs.

9. Discussion

Based on the findings of the assessment, RAG would be most appropriately augmented, but not used to replace enterprise search systems. Key observations include:

- Hybrid retrieval offers a compromise between the accuracy of a more traditional type of search using keywords or other structured search methods and the semantic freedom of embedding-based search.
- The controls required to ensure the safe deployment of production are governance controls, such as access controls, audit logging, and response validation.

The proposed architecture will facilitate the adoption that occurs in phases, such that companies can add the features of RAG to the current search-related infrastructure without significantly disrupting it. However, its success hinges on the factor of quality, which stipulates the continuous enhancement to accommodate new information of the enterprise.

10. Threats to Validity and Limitations

Limited number of limitations need to be mentioned:

- *Generalizability of dataset:* The assessment dataset can be considered a subset of enterprise data (documents, logs, and structured records) and can not be generalized to other spheres.
- *Embedding bias:* The relevance of pre-trained embeddings may drift or become biased over time, affecting retrieval quality [10].
- *Single-tenant focus:* This research focuses on single-tenant deployments, which require more isolation and control over multi-tenant or inter-domain environments.
- *Data dynamism:* Enterprise data are also dynamic, and embeddings and retrieval plans must be updated to prevent adverse impacts on performance.

The limitations in these sections highlight the areas where the findings should be critically read and further tested.

11. Future Work

Future research will focus on improving the adaptability, accuracy, and reliability of the proposed framework. One of the areas that future research will cover is the development of an adaptive retrieval mechanism that incorporates explicit and implicit feedback from users to enhance the precision and accuracy of the proposed framework. Another potential area of research is the domain-conditioned fine-tuning of the embeddings, which will be useful in improving the performance of the proposed framework in handling enterprise domain jargons and schemas. Another potential area that will be explored in the future is the development of an automated prompt optimization mechanism that will be able to optimize and fine-tune the prompts without any manual intervention. Additionally, more tests and experiments will be conducted in the future in various industries and domains to enhance the reliability and generalizability of the proposed framework. Finally, there will be a greater focus on incorporating feedback loops and learning mechanisms in the proposed framework to enhance its reliability and accuracy in the future.

12. Conclusion

This study presents a hybrid enterprise application search framework that integrates sparse and dense retrieval with retrieval-augmented generation (RAG) to address the long-standing limitations of keyword-based enterprise search systems. By positioning large language models as constrained reasoning components layered on top of a governance-aware retrieval pipeline, the proposed architecture improves semantic relevance, contextual reasoning, and operational safety without compromising enterprise requirements, such as access control, auditability, and predictable latency.

The main contributions of this work are as follows: (i) the design of an application-centric hybrid retrieval architecture that combines inverted indexing with embedding-based semantic search; (ii) a controlled RAG reasoning layer that enforces strict grounding and traceability; and (iii) a workflow-oriented evaluation demonstrating the system's applicability to real enterprise tasks, such as incident diagnosis and impact analysis. Quantitative results on real-world enterprise datasets show that the proposed approach improves the top-k retrieval precision and recall by up to 25%, increases task completion accuracy by over 20%, and reduces hallucinated responses by approximately 40% compared to traditional keyword-based search and ungrounded LLM baselines.

These findings indicate that hybrid retrieval-and-reasoning architectures offer a practical and scalable solution for enhancing enterprise application search. Future work will explore adaptive reranking strategies, incremental index updates for highly dynamic artifacts, and broader validation across multi-tenant and cross-domain enterprise environments.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

[1] Dai Z, Yang Z, Yang Y, Carbonell JG, Le QV, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–88, Florence, Italy, 2019 Jul. Available from: https://aclanthology.org/P19-1285.pdf.

[2] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*. Red Hook (NY): Curran Associates; 2020;33:9459–9474.

[3] Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, *et al.* Dense passage retrieval for open-domain question answering. *Proceedings of EMNLP*, pp. 6769–81, 2020 Nov. Available from: https://arxiv.org/pdf/2004.04906v2

[4] Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD. arXiv [Preprint]. 2018. doi: 10.48550/arXiv.1806.03822.

[5] Feldman R, Sanger J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press; 2007.

[6] Mitra B, Craswell N. An introduction to neural information retrieval. *Found Trends Inf Retr*. 2018;13(1):1–126. doi: 10.1561/1500000061.

[7] Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press; 2008.

[8] Chen D, Fisch A, Weston J, Bordes A. Reading Wikipedia to answer open-domain questions. arXiv [Preprint]. 2017. doi: 10.48550/arXiv.1704.00051.

[9] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. *Found Trends Inf Retr*. 2009;3(4):333–89. doi: 10.1561/1500000019.

[10] Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th IJCNLP*, pp. 3816–30, Online, 2021 Aug. doi: 10.18653/v1/2021.acl-long.295.