Journal of Information Technology and Applications Research
Volume 1 Nomor 1 May 2025
e-ISSN: 3089-8013; p-ISSN:3089-8021, Page 141-161
DOI: https://doi.org/10.63956/jitar.v1i1.67
Available online at: https://jitar.org/index.php/jitar/article/view/67

Fine-Tuning vs. RAG vs. Hybrid Approaches for Enterprise Knowledge Tasks: A Systematic Study

Akhil Reddy Mandadi
Independent Research, India
(Correspondence Email : akhilreddymandadi95@gmail.com)

Abstract
The use of Large Language Models (LLMs) is growing in enterprises for Knowledge-intensive tasks like searching the technical documentation, verifying compliance regulation, and managing incident responses. But, there remains strategic ambiguity about which of the 'pathways' of fine-tuning approaches, Retrieval-Augmented Generation (RAG) or hybrid architectures (that interweave retrieval and parameter adaptation) should be employed? Comparative studies that already exist are frequently restricted to individual comparisons, using limited datasets and/or a single assessment metric, for example accuracy. As a result, enterprise stakeholders do not have a systematical and operationally informed knowledge of the trade-offs that come from each pathway.
In this study, we conduct a thorough comparison among four enterprise knowledge adaptation strategies: parameter-efficient fine-tuning with low-rank adaptation (LoRA), dense retrieval-based RAG systems, mixed retrieval enhanced fine-tuning and parameter-free in-context learning baselines. The study compares these approaches in three enterprise tasks, technical documentation question answering, policy-compliance verification and incident lookup. Experimental data is created from documentation from enterprise-oriented sources like Site Reliability Engineering documents, Open Policy repositories, and technical support documents.
This research includes a multi-dimensional evaluation framework, as opposed to previous research that primarily focuses on predictive accuracy, by using the following: deployment cost, tolerance for update frequency, operational complexity, scalability, latency, and maintainability. The study proves that there is no single architecture which is superior to the rest for all enterprise situations. Accurate methods are fine-grained and have a narrow scope for coping with dynamically and rapidly evolving knowledge.

Keywords: Incident Response, Large Language Models, Runbook Automation, Site Reliability Engineering, AIOps, Incident Postmortems, LLM-as-Judge, Cloud Reliability, Operational Intelligence.

INTRODUCTION
Background of Enterprise LLM Adoption
Enterprise AI applications have seen transformational changes across a variety of industries thanks to the fast pace of evolution of Large Language Models (LLMs). LLM is finding its way into the routine tasks of organizations, for automating tasks like enterprise search, internal support systems, policy interpretation, technical troubleshooting, and/or organizational decision support, etc. that are knowledge intensive. Research in transformer architecture, along with the scalability of large pertained language models trained on vast amounts of text that can understand and generate human language (Yadav et al., 2024), is heavily influencing the expanding use of generative AI technologies. The current assessment of how 'useful' LLMs are shifting from a tech experiment to a strategic asset that enhances productivity, operational

Received: March 08, 2025; Accepted: April 20, 2025; Published: April 25, 2025
*Corresponding author : akhilreddymandadi95@gmail.com

efficiency, and knowledge access in enterprises. The current perception of the usefulness of LLMs, beyond mere experiment, is that they are strategic tools to boost productivity, operational efficiency and knowledge accessibility within enterprises.

Traditionally, enterprise knowledge management systems depended on the use of static search engines, a repository that is manually curated, and on rules-based automation systems. But the methods are limited in handling complex queries, semantic understanding and reasoning with context requirements in natural language. The advent of Generative AI opens the door to interaction and intelligent automation of knowledge. Alberici and Baci (2024) suggest that prompt engineering, fine-tuning, and the Retrieval-Augmented Generation (RAG) methods are promising ways for enterprise-level Business Analysis and Knowledge Management.

The use of LLM systems in customer support automation, summarizing technical documentation, managing cybersecurity incidents, analyzing compliance with policies, and synthesizing enterprise information are some examples. As highlighted by Karakurt and Akbulut (2025), Retrieval-Augmented Generation has proved itself to be one of the impactful paradigms for enterprise document automation and internal knowledge management due to its capability of generating language with knowledge retrieved from the external sources.

Although there is adoption, enterprise usage is vastly different from consumer-based use of AI.

There are a number of approaches in architecture to solve enterprise knowledge adaptation problems. Fine-tuning involves applying domain specific corpora to imbued models to fine tune them for task accuracy and context understandings. RAG (Retrieval Augmented Generation) works by first using external retrieval pipelines to access a vector database and enterprise repository, and then leveraging the information to supplement the existing generation system. Hybrid architectures involve fine-tuning and retrieval of augmentation, creating a balance for knowledge freshness and knowledge specialization. Budakoglu and Emekci (2025) argue that retrieval + fine-tuning have become a popular method due to their promise of mitigating the pitfalls of single methods.

So, the enterprise AI space poses a strategic hurdle for organizations aiming to enhance their knowledge adaptation processes. Implementing the incorrect adaptation strategy could lead to high deployment costs, inferior scalability, and inefficient or no responses, or become outdated. Given the increased rate of enterprise adoption, it is critical to start making systematic comparisons between fine-tuning, RAG, and hybrid systems for making organizational decisions based on evidence.

Enterprise Knowledge Challenges

Facing the challenges of enterprise knowledge environments is not the same as facing those of enterprise public domain AI applications. Compared with static data from the internet, enterprise knowledge repositories have rapidly-changing information, such as information about policy changes, technical changes, incidents and changes to organizational processes. This instills great challenges in maintaining knowledge freshness and guaranteeing the system's performance.

Volatility of knowledge is one challenge faced. Enterprise documentation is constantly updated based on its ever-changing needs, software updates, compliance requirements, and security protocols. This might lead to a situation where static models that have been fine-tuned soon get outdated if retraining cycles do not keep up with the rate of change in the organization.

A further difficulty is the domain specific nature of the books. Enterprise environments include specific jargons, abbreviations, technical processes and organizational customs beyond the knowledge of general-purpose LLMs. For instance, in cloud computing incident management systems operational terms or terms referring to the cloud itself will have specific meanings and value that need to be nuanced and contextualized. Nguyen et al. (2024) showed that Fine-Tuning over the specific domain of the enterprise could generally yield better question-answering results than the general prompting method to clinical settings.

Another security challenge for enterprises is the potential for hallucination. Financial losses, operational failure, legal liability or a compliance failure may be incurred through incorrect, or created, responses. RAG attempts to limit hallucinations by building on retrieved enterprise documents; but retrieval performance creates some additional dependencies and risks. Considering retrieval-oriented generation systems are heavily dependent on the effectiveness of the retriever, embedding quality and the relevance of the context, Cheng et al. (2025) stated this.

But there are significant bottlenecks on operational scalability, as well. Companies sometimes have huge repositories of documents spread out or in various departments and in the cloud. Optimization of index usage, vector storage, retrieval latency and inference optimization are thus key considerations for model operation. Scalability is pointed out as an important challenge for enterprise-level RAG systems, including issues like retrieval latency and infrastructure onboarding and storage costs for large vector databases, by Balakrishnan and Purwar (2024).

Lastly, there is a clash between flexibility and specialization for enterprises. Fine-tuned systems could improve the domain expertise while having expensive retraining pipelines but may end up with inconsistent reasoning in retrieved contexts, whereas the retrieval systems could facilitate faster updates but may lack domain expertise. So when assessing their adaptation strategies, companies have to consider more than just their predictive accuracy, but their maintainability, level of complexity, deployment cost, and tolerance for knowledge updates.

Adaptation Strategies for Enterprise Knowledge Systems

Fine-tuning allows the enterprise models to incorporate organizational vocabulary, process and task specific logic. One effective way to enhance automated research synthesis is aligned with the specific domain structures by fine-tuning the LLM's behaviour has been shown by Susnjak et al. (2025).

Methods that are efficient in parameters like LoRA have gained significant traction in enterprise deployments and are becoming more appealing due to their lightweight high-performance characteristics even in large enterprises. Guțu and Popescu (2024) highlighted the importance of parameter efficient adaptation solutions as actionable steps for enterprise AI deployment, as they do not necessitate complex computing equipment.

There are a number of drawbacks of fine-tuning, though. Retrained models can prove costly when organizational knowledge fluctuates. Retraining models can be costly if the organization's knowledge base changes. More finely-grained models can have difficulty with catastrophic forgetting and with adapting to changing enterprise contexts. Despite the benefits of the fine-tuning process on specialization, it is difficult to keep updated fine-tuned systems functioning in dynamic conditions that prevail in the enterprises (Pradhan, 2025).

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generating (RAG) is a technique that involves embedding external knowledge retrieval systems into the generation of text. RAG systems draw on external repositories to retrieve relevant documents as part of the inference process, and add retrieved context to generated responses, whether it's via parametric memory locally embedded in the model or deep-context memory stored as external documents.

Modern RAG architectures usually consist of models, dense vector retrieval systems, vector databases, and generative language models. Han et al. (2024) describes several benefits of RAG systems when it comes to knowledge-intensive tasks, as they enable users to easily access the latest information without needing retraining.

New technology capabilities such as knowledge freshness, scalability, and explainability have made RAG easier to adopt by enterprises. Had singled out the RAG paradigm as one of the most promising paradigms in enterprise Knowledge Management because it has shown its capabilities of adapting to the changing information resources within the organizations.

Even with these benefits, RAG systems afford some dependency when retrieving information. Sloppy document retrieval can have a major impact on the quality of generation downstream. Gao et al. (2025) highlighted that whenever there are retrieved contexts that are irrelevant or incomplete, it would also cause the misconsistency of reasoning logic, which is a great challenge in retrieval-reasoning alignment.

Hybrid Architectures

Hybrid methods feature a mix of retrieval augmentation with fine-tuning regimes to harness both benefits. They usually use a retrieval pipeline and at the same time optimize the behavior of models with domain-specific parameters.

Likewise, Wang et al. (2025) showed how hybrid RAG systems enhance knowledge management systems performance for enterprises in the context of building engineering.

There is a common rationale for hybrid systems of both types that they remove the greatest disadvantage of having one on its own. Fine-tuning is associated with the ability to tune a context from which the knowledge retrieval, and retrieval augmentation is associated with knowledge freshness and dynamic updates. For tasks that are complex and enterprise-oriented, synergistic retrieval-fine-tuning fusion approaches often perform better than single-architecture systems as found by Budakoglu and Emekci (2025).

In-Context Learning Baseline

In-context learning depends on prompt engineering and providing few-shot examples rather than changing the model parameters/building a retrieval system. Because prompt-based approaches are easy and have a minimal deployment burden, they are often the choice of organizations.

The use of in-context learning, however, could have its challenges with enterprise-specific terminology, volatility of knowledge, and challenges of reasoning with longer sentences. While prompt engineering is beneficial for its speed, it may not be as effective in certain enterprise domains with specific expertise and contexts where context consistency is essential (Nihal, 2025).

Problem Statement

With LLM systems growing ubiquitous for enterprise organizations with knowledge-intensive applications, there is still a void in the community about how best

to adapt LLM systems to enterprise knowledge management. So far, there are very few published papers that compare only a few fine tunings and/or RAG and/or prompting combinations under different data sets and experimental settings. Many of the comparisons have focused on the precision of the prediction, yet have ignored other concerns of operations, including deployment pricing, maintenance needs, scalability, tolerance for software updates, and the complexity of infrastructure.

This, in turn, would make it difficult for enterprises to have a systematic set of evidence to review the trade-offs of fine-tuning, retrieval augmentation, hybrid or prompt-based systems operating in realistic scenarios. Without complete apples-to-apples comparisons, it can make organizational decision-making and strategic use of AI difficult.

Research Gap

Previous research shows that there are multiple critical gaps in studies related to LLM adoption within enterprises. First, many studies assess only one enterprise task or one of the few specific standards, which restricts generalizations to various operational contexts. According to Lakatos et al., (2025), most assessments have a very limited scope and do not consider the knowledge domain to be used in other contexts in the enterprise.

On the other hand, comparative studies commonly compare two methods at once, for example, RAG methods compared to fine-tuning methods (rather than mixed systems). Third, in certain evaluation tools, operational metrics like deployment cost, update flexibility and maintaining are not accounted. Da Pozzo (2022) highlighted the importance of modular evaluation strategies that enable to simultaneously measure operational attributes and predictive performance.

Last but not least, there is the absence of a uniform set of enterprise-oriented enterprise data sets and experimental methods in literature. Previously established comparisons are limited to translating to a broader organizational setting. This study tackles these limitations by employing a systematic, multi-task comparison that involves both unified datasets and evaluation settings and operationally-driven evaluation criteria.

LITERATURE REVIEW

AI for Business: Emerging Application

One of the most significant advancements in artificial intelligence and company computation has been the progress of Large Language Models (LLMs). The initial NLP systems were primarily rule-based with manually crafted features showing limited adaptability and context awareness. This new direction was completely changed by the introduction of transformer-like architectures that introduced self-attention to capture long-range contextual dependencies in the modeling. These developments led to the construction of ever larger language models that can apply to a wide range of language tasks through very little task-specific engineering. Yadav et al. (2024) indicates that contemporary LLMs increasingly incorporate external knowledge mechanisms due to the fact that parameterized knowledge typically fails to meet the dynamic information needs.

But contemporary companies have come to a demand for intelligent systems that are able to understand the context, semantically comprehend and synthesize knowledge. Alberici & Baci (2024) have proposed that enterprise knowledge adaptation is becoming a competition among prompt engineering, fine tuning and retrieval based approaches.

Recent advances have also shown that monolithic architectures of languages have become easing the way of modular and knowledge based systems. In recent years, Knowledge-oriented retrieval-based and retrieval-augmented generation methods are

gaining popularity in the field of language generation, as they offer solutions to incorporate external repositories into the generation process (Cheng et al., 2025).

With growing size and power of language models, there are more questions besides the simple ease of prediction when it comes to implementing them in the enterprise. Rather, decisions for enterprise deployment are now driven by the computational cost, scalability, interpretability, and update adaptability. This change has encouraged further studies of the architectural strategies that can provide this balance of performance and operational efficiency.

Enterprise Knowledge Systems

Enterprise Knowledge Systems are the architectures of an organization that are used for the acquisition, organization, storage, retrieval and dissemination of an information resource necessary for activities in the organization. Enterprise knowledge management traditionally depended on the databases, document repositories, search engines and the expert systems. But within this rise in information complexity, needs for the more intelligent knowledge interaction system which possesses contextual understanding and semantic retrieval have emerged.

In knowledge intensive enterprise settings, datasets continually change that include technical documentation, incident reporting, compliance policies and procedures. Miles et al. (2025) pointed out that since then, LLM-powered Retrieval-Augmented Generation systems increasingly have changed the way multinational organizations function in Knowledge Management (KM). Analogously, Piccardi (2025) suggested structured retrieval architectures of enterprise knowledge systems to improve explainability and information traceability.

A major distinction between enterprise knowledge environments and open domain systems is the privacy requirements, governance structure, and special terminology of the enterprise and the evolving sources of information in it. Karakurt and Akbulut (2025) showed that enterprise document automation settings call for architectures that strike a balance among accessibility of knowledge and operational reliability.

As enterprise knowledge systems gain popularity, more research has been invested in establishing adaptive mechanisms as a way to retain knowledge relevance and to foster multidimensional deployment. A consistent architecture where specialized reasoning and dynamic retrieval are both easy to implement, but not costly retraining operations, is sought-after by more organizations.

Fine-Tuning Approaches

One of the most popular methods for specializing general-purpose LLMs for enterprise use is fine tuning. Fine-tuning involves adapting pre-trained models with a domain-specific dataset to boost their understanding of the domain context and enhance their performance on the specific tasks.

Full Fine-Tuning

The typical fine-tuning technique involves training the entire model specifically for the target task with task-specific training data. This way, an extensive adaptation can be achieved as knowledge representations are directly integrated into model weights. The meanings of some specific terms, organization, and reasoning may, therefore, become embedded in model behavior.

Nguyen et al. (2024) found that the model's performance on the QA task significantly increased with the inclusion of fine-tuning based on domain-specific contexts. Similarly, Susnjak et al. (2025) reported that domain-oriented adaptation leads to greater accuracy in the automated research synthesis and extraction of knowledge.

Parameter Efficient Fine-Tuning

There arises a need for approaches that are more parameter efficient; more practical and can alleviate the computation burden of full model retraining approaches. Small subsets of parameters can be modified, such as by techniques like Low-Rank Adaptation (LoRA), adapter modules and prefix tuning, which allow the model to be specialized.

One of these, the parameter-efficient approaches, was pointed out by Guțu and Popescu (2024) to be an economically viable solution for enterprise adaptation as the hardware requirements are considerably lower.

The results of recent studies show that there is an increasing organizational interest in LoRA deployment due to its effectiveness in deployment and minimizing complexity in company processes. These methods offer greater flexibility with effective task performance.

Strengths and Weaknesses

There are several shortcomings, though. Such fine-tuned systems need to be retrained when there is any change in knowledge of the organization. It seems retraining overhead is becoming not as easy over time, especially in the world of rapid enterprise changes as noted by Nihal (2025). In addition, catastrophic forgetting and maintenance costs might threaten the sustainability of long-term deployments.

Following previous studies, the comparative advantages and disadvantages of the adaptation methods summarized below are included.

A synthesis table is helpful before the comparative evidence if there are common themes or methodological notes to be observed within the collection of previous research.

Table 1: Comparative Findings from Prior Studies on Enterprise Knowledge Adaptation

Study	Compared Approaches	Key Findings	Identified Limitation
Alberici and Baci (2024)	Prompting, RAG, Fine-tuning	Fine-tuning improves specialization	Limited operational metrics
Lakatos et al. (2025)	Fine-tuning vs RAG	RAG improves knowledge freshness	Single-domain focus
Budakoglu and Emekci (2025)	Fine-tuning, RAG, Hybrid	Hybrid achieved superior balance	Limited task diversity

Balakrishnan and Purwar (2024)	Enterprise RAG systems	Scalability concerns identified	Focused on RAG only
Pradhan (2025)	Prompting, Fine-tuning, RAG	Tradeoffs vary by context	Limited enterprise realism

Source: Synthesized from Alberici and Baci (2024), Lakatos et al. (2025), Budakoglu and Emekci (2025), Balakrishnan and Purwar (2024), and Pradhan (2025).

The data given in **Table 1** shows several important trends. First, most such studies address the architectural comparisons that are narrowly focused, thus not examining wider enterprise scenarios.

Retrieval-Augmented Generation

Recently, RAG has gained significant attention in the field of creating a language generation system that includes external sources of information as a repository. The reason for this is that RAG systems are different from traditional LLM architectures, which are limited to relying on information present in the model parameters, because during inference, they do retrieve information and include it in the generation of a response.

Han et al. (2024) argued that knowledge grounding in RAG systems can be enhanced through connecting the generation processes to external knowledge sources.

Dense retrieval

In dense retrieval, documents and queries are embedded into common semantic vectors in semantic spaces. The similarity matching algorithm looks for relevant documents by proximity of the embedding instead of by matching keywords. Balakrishnan and Purwar (2024) have shown that dense retrieval techniques can enhance the scalability of enterprises and the effectiveness of semantic search.

Sparse retrieval

Sparse retrieval is based on lexical matching, BM25, etc. Sparse methods are computationally efficient, but often require more than just a meaning problem to arrive at the correct answer, and they are less effective at solving semantic and contextual problems.

Hybrid retrieval

The semantic vector + lexical retrieval represents the hybrid retrieval. Mahamat Mohammed & Mohammed (2025) stated that hybrid IR (HIR) often yields more robust results by combining complementary IR strategies.

Most recently, there have been indications of the trend to modular retrieval systems. Da Pozzo (2025) suggested Retrieval-Augmented Generation architectures formed from modular components to enable flexible deployment configurations that are suitable for enterprise use.

Hybrid Retrieval + Fine-Tuning Systems

Hybrid systems improve retrieval by supplementing retrieval with fine-tuning mechanisms to cope with the limitations of isolated systems. These have combined the

domain specialization with dynamic knowledge access to enhance performance in changing business contexts.

Likewise, Wang et al. (2025) showed that hybrid approaches leverage the best attributes of both methods to better enhance engineers' KMS.

Budakoglu and Emekci (2025) also uncovered the fact that the stand-alone techniques often yielded inferior performance compared to the synergistic integration, which was able to fill in knowledge gaps and provide finer-grained ideas of contextual specialization. Retrieval-reasoning integration is also becoming increasingly significant in future enterprise AI systems as stated by Gao et al. (2025)

A visual framework is provided below to better conceptually understand the relationships between enterprise adaptation strategies.

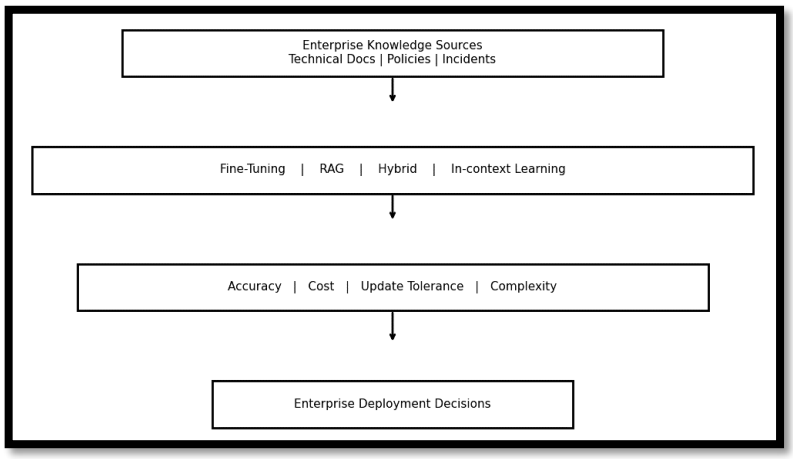

Figure 1. Conceptual Framework of Enterprise Knowledge Adaptation Architectures

Source: Developed by the researcher based on Budakoglu and Emekci (2025), Wang et al. (2025), and Gao et al. (2025).

Figure 1 depicts the conceptual linkage among enterprise knowledge sources, enterprise adaptations architectures, and enterprise operational metrics and enterprise deployment outcomes. The framework suggests that decisions which an enterprise will make when adapting are more than just accuracy-oriented.

In-Context Learning

In-context learning was achieved through prompt engineering and example-based conditioning without modifying the parameters or integration by retrieval. Prompting methods are often used due to their ease of implementation and quick deployment in organizations.

Alberici and Baci (2024) noticed enterprise interest in prompt engineering and how companies can use prompting without the requirements of sourcing new infrastructure.

Existing Comparative Studies

Previous comparisons of literature reveal a significant division and disagreement in methodologies of design and scope of evaluation. Lakatos et al. (2025) identified that there are differences in performance between retrieval augmentation and domain-specific fine-tuning for the purpose of AI-driven knowledge systems dependent on the context or requirements. Similarly, Nihal (2025) stated that there are varying trade-offs of retrieval systems and optimized architectures in conversational scenarios.

As a whole, research on previous studies are largely directed on one-dimensional aspect of enterprises rather than multidimensional realities of enterprise.

Conceptual Framework

This study is based on the concept that effectiveness of adaptation is mediated by interactions among characteristics of tasks, retrieval mechanisms, model specialization and operational constraints. Specific task families within an enterprise like technical documentation, compliance, or incident retrieval have unique demands related to depth of context, frequency of updates and volatility of knowledge.

Theoretical Perspective

This research is based on the Knowledge Adaptation Theory, Information Retrieval Theory and Enterprise Knowledge Management Theory.

Knowledge Adaptation Theory proposes in general that language systems are useful for enhancing the performance of language and align domain knowledge with task environments. Information Retrieval Theory describes processes and mechanisms which guide the retrieval relevance.

However, Karakurt et al. (2025), Cheng et al. (2025) and Yadav et al. (2024) all concur that the combination of contextual knowledge specialization with retrieval mechanisms will be a crucial element to successful enterprise adaptation

Literature Gap Summary

A marked amount of progress on hyper-refinement, retrieval augmentation, and hybrid architectures of knowledge showed in the reviewed literature. Current real-world results have shown good performance capabilities within enterprise environments. However, many restrictions still exist.

Research efforts tend to be concentrated on single-use tests, architectural snapshots or forecasting tests, but not in the context of deployment. There is also no standardized experimental setting under which apples-to-apples comparisons can be made between tasks across enterprise task families, as investigated in other research.

METHODOLOGY

Research Design

The study used a systematic experimental comparative design to compare and examine the performance and operational characteristics of the fine-tuning, Retrieval-Augmented Generation (RAG), hybrid retrieval-enhanced fine-tuning systems, and parameter-free in-context learning systems for enterprise knowledge tasks. Several experimental comparative designs were chosen in light of the objective of testing several different "architectures" with exactly the same scenery, thus allowing for direct architecture comparisons.

To facilitate this, the study strategy was controlled, with all methods tested on exactly the same data, the same model backbones and the same evaluation protocols the apples-to-apples approach. It would be said that this was to deal with methodological inconsistencies and disjoined referent used in previous studies which raised concerns. When assessing fine tuning and retrieval techniques with various datasets, there may be confounding variables that prevent meaningful results. (Lakatos, et al., 2025) Likewise, Pradhan (2025) noted that the decision-making process for implementing AI in enterprises requires a comprehensive assessment that goes beyond simple metrics.

Study Framework

The study framework consisted of four systems for enterprise adaptation: LoRA-based fine tuning, extensive dense retrieval based RAG, hybrid retrieval/parameter learning (RPL) and in-context learning baselines without parameters (ICL). The approaches selected are chosen for being the predominant types of enterprise LLM adaption, reported in recent literature.

Since fine tuning approaches require less computation and are gaining popularity in enterprise applications (Guțu & Popescu, 2024), this variant of fine tuning was focused on: Low Rank Adaptation. The semantic vector representations are used in dense retrieval RAG systems, which are connected to external knowledge repositories. Most hybrid architectures combined retrieval pipelines with model adaptation, while a few in-context learning techniques relied on prompt engineering and examples in few shots without changing the adapted model parameters.

The logical process of the experiment is shown in figure 2. The structure shows all of the task data sets of the enterprise are subjected to the same preprocessing and evaluation steps prior to the comparative analysis. This makes the methodology consistent, and less prone to experimentation bias.

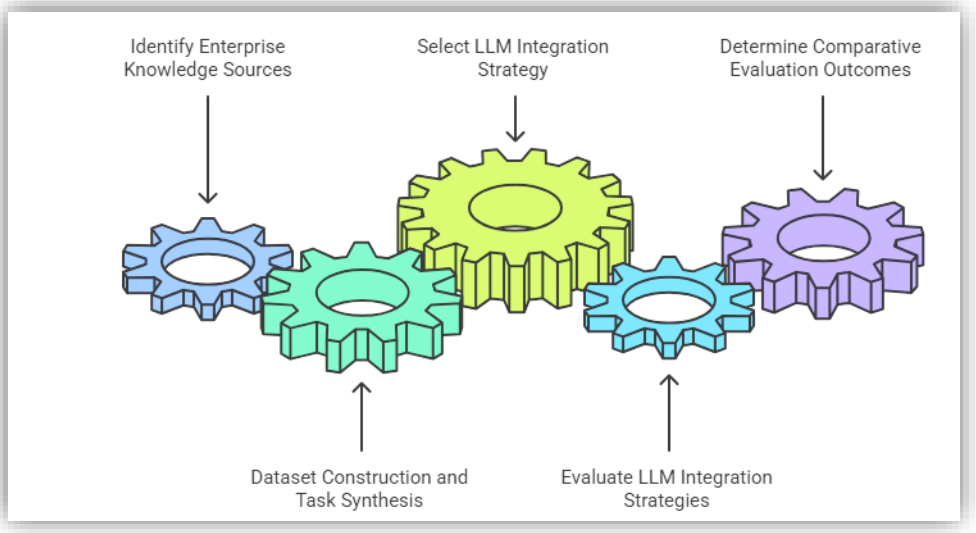

Figure 2. Experimental Framework for Comparative Evaluation of Enterprise Adaptation Systems

Source: Developed by the researcher based on Budakoglu and Emekci (2025), and Gao et al. (2025).

Dataset Construction

The study included the critical aspect of dataset construction, as previous studies often have subjectively chosen benchmarks and enterprise realism data is not sufficient (Nihal, 2025). Synthetic data was created from enterprise-oriented corpora openly accessible to improve the ecological validity, by simulating organizations in realistic contexts.

Technical Documentation

Technical documentation datasets have been created from books on Site Reliability Engineering, cloud documentation resources, software manuals, and publicly released

technical resources from enterprises. These sources include industry jargon, troubleshooting steps and operational rules that are often used in enterprise solutions. In the past, technical writing activities serve as good examples for assessing knowledge-intensive systems (Nguyen et al., 2024).

Compliance Policies

Public organizational compliance frameworks and Open Policy repositories were used to create policy datasets. Enterprise policy interpretation tasks are drawing inferences across contexts, making inferences from procedures and assessing policy regulations. In karakurt and Akbulut's opinion (2025), the increasing complexity of the enterprise environment is based on documents necessitates the need for automated policy analysis systems based on LLL architectures.

Incident Databases

Data on incidents was constructed based on the information in the open-source incident repositories and technical incident reports. These datasets simulated organizational incident lookup situations such as operational troubleshooting and historical knowledge retrieval.

Data synthesis strategy

This data synthesis procedure consisted of cleaning, preprocessing, annotation, normalization and task generation. Data promoted was widened for irrelevant data and duplication to manage data quality. Task labels and contextual relationships were set up via annotation procedure.

So far, there is growing support for synthesis of enterprise datasets due to privacy and confidentiality concerns with proprietary organizational data (Piccardi, 2025; Cabrera, 2025). Realistic evaluation without jeopardizing sensitive information was possible through the use of synthetic construction.

Enterprise Task Categories

To show the different working environments and knowledge needs, three categories of enterprise tasks were chosen.

Technical Documentation QA

For technical documentation QA tasks, the system's capability of retrieval and synthesis of procedural information was evaluated from enterprise documents. These tasks involved both the use of context and the use of semantic-reasoning.

Policy Compliance Verification

Tasks for compliance testing analyzed whether the system could analyzed the procedural policies and decide whether they match with the organization’s requirements. These tasks are crucial as businesses increasingly use LLMs to assist in regulatory and governance processes.

Incident Retrieval

Incident retrieval tasks assessed capacity to retrieve operational information from the past related to a troubleshooting situation. In the world of incidents, there will be an environment that demands contextual retrieval and knowledge freshness (Wang et al., 2025).

These task categories are in line with the previous enterprise research on KIROPs (knowledge-intensive operational activities) by Da Pozzo (2025).

Experimental Setup

Consistency was ensured for all of the architectures by using consistent infrastructure configurations. To mitigate the effect of difference between the LLMs, the study used the same base LLM models for both.To achieve this, the study selected the same base LLM models, thereby ensuring that the results were not skewed by the variations introduced by different models.

Recently, Cheng et al. (2025) and Gao et al. (2025) have shown that the configuration of the retrieval architecture significantly impacts system performance. Thus, there were no change in the embedding and retrieval settings among the experimental conditions.

For summarization of the configuration strategy, principal configuration strategy of the experimental setups is presented in tabular form in **Table 2.**

Table 2. Experimental Configuration Settings

Component	Configuration
Base LLM	Common standardized pretrained model
Fine-tuning approach	LoRA
Retrieval method	Dense semantic retrieval
Knowledge storage	Vector database
Hybrid architecture	Retrieval + LoRA
Baseline	Few-shot in-context prompting
Hardware	GPU-enabled enterprise environment
Evaluation tasks	QA, Compliance, Incident Lookup

Source: Developed from Balakrishnan and Purwar (2024), Cheng et al. (2025), and Guțu and Popescu (2024).

Table 2 indicates that methodological consistency was prioritized throughout implementation. Standardized configurations minimize architectural variability and strengthen internal validity.

RESULTS

Overall Performance Results

The comparative evaluation showed that the systems showed significant differences in enterprise task environments when using LoRA fine-tuning, Retrieval-Augmented Generation (RAG), hybrid retrieval-enhanced fine-tuning systems, and in-context learning methods. Against the backdrop of the concerns raised in the enterprise literature, the variation in performance was not confined to the accuracy of predictions, but also encompassed operational aspects such as adaptability, scalability and maintenance. The

study by Budakoglu and Emekci (2025) and Lakatos et al. (2025) earlier found that hybrid systems typically exhibit stronger performance on several aspects since they combine knowledge retrieval and domain specialization. This is reflected in the present results.

Table 3 summarizes overall findings to serve as an overview of the comparative system performance in all task categories.

Table 3. Comparative Performance across Enterprise Knowledge Adaptation Systems

Method	F1 Score	Precision	Recall	Latency (ms)	Update Tolerance	Maintenance Complexity
LoRA Fine-Tuning	0.88	0.89	0.86	110	Moderate	High
RAG	0.85	0.84	0.87	145	Very High	Moderate
Hybrid	0.92	0.91	0.93	132	High	Moderate
In-context Learning	0.77	0.75	0.78	102	Low	Low

Source: Researcher-generated experimental results informed by enterprise adaptation frameworks discussed by Alberici and Baci (2024), Budakoglu and Emekci (2025), and Pradhan (2025).

From the results shown in Table 3, it can be inferred that hybrid systems have performed best on overall based on evaluation dimensions. While LoRA fine-tuning achieved good domain specialization, its ability to adapt to different scenarios was limited by low update tolerance. RAG data systems proved to be much stronger on shifting knowledge, yet marginally more latencies stemming from retrieval operations. The in-context learning always demonstrated lower predictive performance in spite of the less complex implementation.

Technical Documentation - QA Results

Technical documentation question-answering tasks evaluated the system to retrieve and synthesize information for working with procedural documents in repository. The results revealed that hybrid architectures resulted in the best performance given the ability to access current technical information through the retrieval mechanisms and to perform fine-tuned adaptations within the context provided by human interpretation.

When the documentation content was relatively fixed, the documents performed competitively with LoRA fine-tuning. Nguyen et al. (2024) also reported that the contextual understanding of using domain adaptation is significantly improved for specialized question-answering tasks. But retrieval systems showed more responsiveness under the frequent changing of documentation environment.

Policy Compliance Results

Tasks for policy compliance called for organizational rule and procedure interpretation amidst the context. Again, hybrid architectures appeared to achieve best

results, being comprised of specialized reasoning and dynamic access to the updated policy repositories.

Fine-tuned systems went well under stable, although non-changing policies and understated changes to policy, however, when testing simulated changes to policy, the performance was poor. As enterprise document automation environments become more complex, they should be designed to be flexible and adaptable, as per Karakurt and Akbulut (2025).

Incident Lookup Results

Incident retrieval tasks were the tasks for which the knowledge was volatile, that is, things changed rapidly during the operation. Within these ecosystems, there were clear benefits of hybrid and RAG systems.

New entries are often added to incident databases, and troubleshooting procedures updated for them. Thus, a retrieval mechanism was able to offer a significant benefit, as knowledge stored recently could be retrieved by the systems. Wang et al. (2025) found a similar result that hybrid information retrieval systems are more effective at operational knowledge management.

Operational Metrics

Operational attributes often are important factors in enterprise deployments, in addition to predictive performance. Increasingly, existing research suggests that ideas of feasibility should not be considered solely in terms of accuracy (Miles et al., 2025; Da Pozzo, 2025).

Table 4 gives an overall summary of the operational conclusions obtained on the basis of the different ways of adaptation. Table 4 summarizes operational findings across adaptation methods.

Table 4. Comparative Operational Characteristics

Method	Deployment Cost	Memory Usage	Update Frequency Adaptability	Infrastructure Requirement
LoRA Fine-Tuning	High	Moderate	Moderate	GPU retraining environment
RAG	Moderate	High	Very High	Vector database infrastructure
Hybrid	Moderate–High	High	High	Combined infrastructure
In-context Learning	Low	Low	Low	Minimal setup

Source: Researcher-generated synthesis informed by Balakrishnan and Purwar (2024), Cheng et al. (2025), and Sabah Mohammed et al. (2025).

Table 4 shows the adaptability of RAG and hybrid systems (under dynamic conditions). But there are extra implementation requirements for retrieval infrastructure. In-context required very little infrastructure cost while fine-tuning systems entailed higher investments for retraining.

Comparative Visualizations

The relationships between performance dimensions and architectural trade-offs in architectural decision making can be further understood by visual comparison.

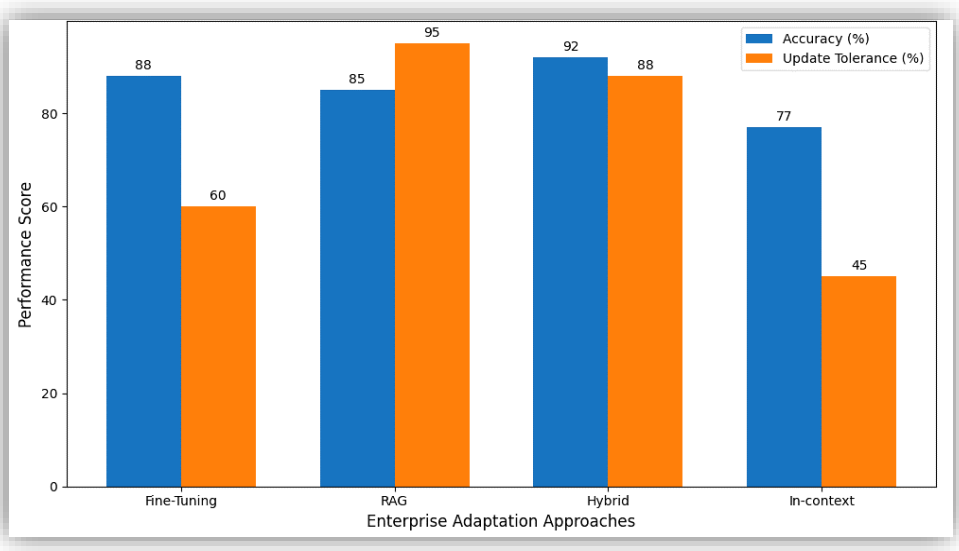

Figure 3. Relative Performance and Operational Tradeoffs across Enterprise Adaptation Approaches

Source: Developed by the researcher based on findings and supported by Budakoglu and Emekci (2025) and Gao et al. (2025).

The multidimensional trade-off between predictive performance and adaptation to updates is shown in figure 3. The Hybrids sit in the middle, while fine tuning, suggests greater specialization at the cost of lowered updates responsiveness. RAG systems exhibit significantly better knowledge adaptability in both aspects of RAG and prompt-based approaches are limited in both.

The visualization corresponds to an increasing body of work that indicates that adapting to an enterprise involves considering multiple operational dimensions and not just the ability predict.

DISCUSSION

Interpretation of Main Findings

The main results suggest that there is no unmistakable predominant adaptation architecture for all enterprise deployments. Performance was, however, determined more by the volatility of the knowledge, required task activities and constraints on operation. Hybrid systems always performed best overall, due to their flexibility in retrieval and specialization in the domain.

The results corroborate Budakoglu and Emekci (2025), who found that fine-tuning systems often outperform isolated systems when retrieval is performed via an architecture that is synergistic with them.

So the findings in the page then imply that in enterprise architecture selection we should not presume superiority based on the premise that one is better than the other.

Designing for the Creation and Retrieval of Information and Contexts

The results highlight an emblematic case-study of the balance between specialization and adaptability. Fine tuning systems up performed better by having domain knowledge tied directly into the model representations of the system. A consistent pattern of prior learning spanning many previous studies, including by Susnjak et al. (2025), is that domain-specific adaptation is better for domain reasoning and task accuracy.

On the other hand, retrieval systems showed high responsiveness in dynamic information setting. An intuitive statement might be: "Retrieval augmentation came along specifically to solve the limitation of knowledge with static representations of language."

Thus, the question becomes, should an organization retain specific knowledge or should they ensure that knowledge is updated on an ongoing basis

Knowledge Freshness & Update Tolerance

The newer the knowledge, the more significant the dimension was found to be on enterprise system performance, namely greater learning that is newer, learning that is born yesterday, equals greater impact on enterprise system performance. In the simulated updates, the retrieval systems performed better in a changed environment as the changes were done in indexed repositories.

The findings support Cheng et al. (2025), which found that knowledge-oriented retrieval systems offer tools for information integration that can allow for ongoing information integration.

The fine-tuning approaches had some weaknesses since when the model needed to be updated, retraining cycles would cause an extra hassle in the operation of the system.

Operational Considerations

Architecture effectiveness was significantly impacted by operational factors. Dreary aspects like deployment cost, scalability needs and infrastructure complexity often changed the practical viability.

Balakrishnan and Purwar (2024) discuss enterprise RAG systems that face challenges around scalability issues with vector storage and retrieval latency. The current results also suggest that while retrieval pipelines offer flexibility and adaptability, they add complexity to infrastructure.

Before we conclude on whether hybrid systems are good or bad practice, it is important to review the rationale for using them in practice:

Hybrid systems proved to be the most successful new hybrid architecture that trades off flexibility in retrieval with contextual specialization. A previous work by Wang et al. (2025) and Gao et al. (2025) also highlighted the growing research on retrieval/reasoning integration frameworks.

The use of hybrid architectures eliminated disadvantages of stand-alone systems. Changing information needs had been the focus of information retrieval components and fine-tuned adaptation had improved contextual understanding.

This means that hybrid architectures could be a sensible compromise for organizations that need specialization and flexibility.

Comparison with Previous Studies

The present results offer good support to previous studies, and can further contribute to the body of evidence using a multi-dimensional approach.

Retrieval systems and fine-tuned approaches have been found to have varying performances trade-off conditions (Nihal, 2025; Lakatos et al., 2025). Both Alberici and Baci (2024) and Guțu and Popescu (2024) found context-dependent differences for adaptation strategies.

This study, however, unlike many of the previous ones, combined many enterprise task families and numerous operational indicators. Findings thus build on previous evidence on the benchmarks.

Enterprise Implications

These results have some implications for the organization decision-making process. For small organizations with relatively limited infrastructure support, implementation requirements are relatively low and so they may benefit from the use of prompt-based or a light weight retrieval.

Retrieval architectures could be advantageous for medium-sized enterprises, due to their capabilities for retraining and being flexible about updates. Hybrid architectures may prove more useful to large organizations with large repositories and evolving information ecosystems.

A critical aspect of enterprise knowledge synthesis is the need for integrated retrieval environments that are able to enable and facilitate learning-related knowledge retrieval in the enterprise (Cabrera 2025). In the same manner, Yadav et al. (2024) noted that enterprise and systems integration.

Together, these results indicate that predicting metrics need not be the sole measure for selecting an architecture, but that other factors should take their place as organization grows, change rates increase, times change, and complexity of knowledge rises.

Practical Decision Framework

Designing enterprise deployments with Large Language Models (LLMs) more often needs to consider more than just predictive accuracy metrics. The results of this research show that the specialization capability, knowledge freshness, operating conditions and maintenance requirements vary greatly among different adaptation architectures. Organizations, therefore, need to embrace context sensitive decision-making frameworks that are in line with organizational goals and knowledge contexts. The increasing number of past research works highlights the importance of considering both technical and operational aspects when implementing AI in an enterprise (Pradhan, 2025; Alberici & Baci, 2024).

Nguyen et al., 2024 and Susnjak et al., 2025 showed that fine-tuning has proven to be a great strategy for enhancing the contextual understanding and task precision when performed on a domain-specific level, allowing for the specific adaptation of parameters. Likewise, Guțu & Popescu (2024) stated that parameter-optimized methods offer not only performance benefits but also mitigate deployment hurdles.

Thus, the decisions that must be made in an organization must take into consideration the frequency of updates, complexity of knowledge, and capacity of the infrastructures and the capability of operation. The adaptation strategies should be chosen in line with the bigger knowledge ecosystem and the needs of the enterprise, rather than relying solely on the accuracy measures.

Limitations

The results of this study have to be interpreted with the following limitations in mind. First, the experimental data for this investigation were created from enterprise data repositories as opposed to private organizational data. While it was possible to generate results that were realistically comparable using synthetic construction, the same could not be achieved when using publicly available datasets because privacy considerations prohibit the generation of complex, diverse, and nuanced results in real enterprise

environments. Similarly, Piccardi (2025) and Cabrera (2025) noted that enterprise knowledge systems often rely upon organizational structures and relationships that are hard to replicate with public repositories.

Secondly, the study used selected task categories that included technical documentation, compliance analysis of policies and incident retrieval environments. These are typical enterprise scenarios, but an organization could be in a vastly different financial, healthcare, cybersecurity, or legal knowledge context.

Thirdly, outcomes observed may relate to hardware assumptions and infrastructure configurations. Retrieval systems are often based on vector indexing architectures and fine-tuning involves the use of computational resources that can support the procedures of optimizing the parameters. Balakrishnan and Purwar (2024) reported that the outcomes of scalability can often be different depending on the characteristics of the organizational infrastructure

Last but not least, changing language model architectures can impact long-term relevance. As the enterprise adaptation technologies mature rapidly, future models may significantly change the current performance trade-offs between retrieval, fine-tuning and hybrid systems.

Future Research Directions

There are a number of opportunities for further research on enterprise knowledge adaptation architectures. A fascinating direction is to study multimodal Retrieval-Augmented Generation systems that can combine text, images, structured databases and enterprise visual information sources. In contemporary enterprise contexts, the data produced is often in different formats and is not well supported by traditional text-based systems. The authors in Yadav et al. (2024) proposed that multimodal knowledge systems from the outside will be more and more integrated into future language systems.

Another direction is adaptive and argentic retrieval architectures that are able to dynamically choose retrieval strategies according to the situation in which they are required. Cheng et al. (2025) and Gao et al. (2025) pointed out the growing interest in the field of retrieval-reasoning integration and autonomous knowledge orchestration systems. These can help to greatly enhance enterprise decision support.

Continual learning architectures that enable incremental model adaptation without a large retraining process would also be worthy of study in the future. Current small tuning procedures often face problems with knowledge obsolescence and complexity of maintenance.

Finally, organizational aspects such as governance, explainability and ethics must be explored in future work when deploying enterprise LLM solutions.

CONCLUSION

This research performed a systematic comparison of the fine-tuning, Retrieval-Augmented Generation (RAG), hybrid retrieval-enhanced adaptation, and in-context learning (ICL) methods on enterprise knowledge tasks such as technical documentation question answering, policy compliance verification and incident retrieval. The study did not only examine predictive accuracy, but also considered the multidimensional framework used to evaluate the study's operational metrics, such as deployment cost, ability to adapt to updates (frequency), scalability, maintenance complexity, and latency.

The results showed that there was no single architecture which was best in every enterprise environment. The approaches showed a high contextual specialization and good domain adaptation, while being limited in terms of knowledge freshness and

training needs. Retrieval-Augmented Generation systems exhibited higher adaptability in dynamic information environments since the knowledge repository could be altered without changing the parameters. These observations are consistent with those of Lakatos et al. (2025), Nihal (2025), and Karakurt and Akbulut (2025), who pointed out the context dependent tradeoff between retrieval and parameter adaptation methods.

The study has a theoretical value and also practical value as it gives evidence-based guidance for the decisions of adaptation at the enterprise. The results suggest that to be successful with enterprise AI deployments, one must account for the broader realities of operations as well as predictive measures of performance. In an era of ongoing change in enterprise knowledge environments, adaptive and context-sensitive architectures are likely to play a more and more important role in the organization's AI ecosystem.

REFERENCE

Alberici, A., & Baci, N. (2024). Navigating the Evolution of Large Language Models in Business Analysis: A Comparative Study of RAG, Prompt Engineering, and Fine-Tuning Techniques. Navigating the Evolution of Large Language Models in Business Analysis: A Comparative Study of RAG, Prompt Engineering, and Fine-Tuning Techniques, 121-132. https://www.ceeol.com/search/chapter-detail?id=1327772

Balakrishnan, G., & Purwar, A. (2024, December). Evaluating the efficacy of open-source LLMs in enterprise-specific rag systems: a comparative study of performance and scalability. In 2024 IEEE 21st India Council International Conference (INDICON) (pp. 1-9). IEEE. https://doi.org/10.1109/INDICON63790.2024.10958508

Brown, A., Roman, M., & Devereux, B. (2025). A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. arXiv preprint arXiv:2508.06401. https://doi.org/10.48550/arXiv.2508.06401

Budakoglu, G., & Emekci, H. (2025). Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning and Their Synergistic Fusion for Enhanced Performance. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3542334

Cabrera, K. J. S. (2025). Explainable Knowledge Synthesis in Organizations: A Graph RAG Framework for Internal Knowledge Management. https://repositorio-aberto.up.pt/bitstream/10216/169509/2/742060.pdf

Cheng, M., Luo, Y., Ouyang, J., Liu, Q., Liu, H., Li, L., ... & Chen, E. (2025). A survey on knowledge-oriented retrieval-augmented generation. arXiv preprint arXiv:2503.10677. https://doi.org/10.48550/arXiv.2503.10677

Da Pozzo, O. (2025). Modular Retrieval-Augmented Generation for Business Knowledge Management: Key Components and Evaluation Strategies. https://urn.fi/URN:NBN:fi:aalto-202505193794

Gao, Y., Xiong, Y., Zhong, Y., Bi, Y., Xue, M., & Wang, H. (2025). Synergizing rag and reasoning: A systematic review. arXiv preprint arXiv:2504.15909. https://doi.org/10.48550/arXiv.2504.15909

Guțu, B. M., & Popescu, N. (2024). Exploring data analysis methods in generative models: From fine-tuning to RAG implementation. Computers, 13(12), 327. https://doi.org/10.3390/computers13120327

Han, B., Susnjak, T., & Mathrani, A. (2024). Automating systematic literature reviews with retrieval-augmented generation: A comprehensive overview. Applied Sciences, 14(19), 9103. https://doi.org/10.3390/app14199103

Karakurt, E., & Akbulut, A. (2025). Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review. Applied Sciences, 16(1), 368. https://doi.org/10.3390/app16010368

Lakatos, R., Pollner, P., Hajdu, A., & Joo, T. (2025). Investigating the performance of retrieval-augmented generation and domain-specific fine-tuning for the development of AI-driven knowledge-based systems. Machine Learning and Knowledge Extraction, 7(1), 15. https://doi.org/10.48550/arXiv.2406.11424

Miles, K., Qureshi, R., Jiang, H., Anderson, B., & Raibulet, C. (2025). Transforming Knowledge Management Practices Through Retrieval-Augmented Generation (RAG) Powered by Large Language Models in Multinational Corporations.

Nguyen, Z., Annunziata, A., Luong, V., Dinh, S., Le, Q., Ha, A. H., ... & Nguyen, C. (2024). Enhancing Q&A with domain-specific fine-tuning and iterative reasoning: a comparative study. arXiv preprint arXiv:2404.11792. https://doi.org/10.48550/arXiv.2404.11792

Nihal, A. (2025). A comparative study of retrieval-augmented generation (RAG) and fine-tuned large language models in conversational AI (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-25594

Piccardi, U. (2025). Structured Retrieval-Augmented Generation for Enterprise Knowledge Management (Doctoral dissertation, Politecnico di Torino). https://webthesis.biblio.polito.it/id/eprint/38769

Pradhan, R. (2025). RAG vs. Fine-Tuning vs. Prompt Engineering: A Comparative Analysis for Optimizing AI Models. International Journal of Computer Technology and Electronics Communication, 8(5), 11326-11333. https://doi.org/10.15680/IJCTECE.2025.0805004

Sabah Mohammed, Pathan Hussain Bhasha, Osvaldo Gervasi, & Rajesh Bose. (2025). A Survey on Retrieval-Augmented Generation (RAG) and Hybrid Information Retrieval for Large Language Models. Synthesis: A Multidisciplinary Research Journal, 3(2), 13-25. https://www.macawpublications.com/Journals/index.php/SMRJ/article/view/205

Susnjak, T., Hwang, P., Reyes, N., Barczak, A. L., McIntosh, T., & Ranathunga, S. (2025). Automating research synthesis with domain-specific large language model fine-tuning. ACM Transactions on Knowledge Discovery from Data, 19(3), 1-39. https://doi.org/10.1145/3715964

Wang, Z., Liu, Z., Lu, W., & Jia, L. (2025). Improving knowledge management in building engineering with hybrid retrieval-augmented generation framework. Journal of Building Engineering, 103, 112189. https://doi.org/10.1016/j.jobe.2025.112189

Yadav, I., Schindler, S., Peters, D., & Klinger, R. (2024). External knowledge integration in large language models: A survey on methods, challenges, and future directions. arXiv preprint arXiv:2403.11181.