Volume 12, No. 11, November 2025

Journal of Global Research in Mathematical Archives

ISSN 2320 – 5822

RESEARCH PAPER

Available online at http://www.jgrma.com

AI IN CLOUD-BASED INFORMATION RETRIEVAL AN EXPLORATION OF CURRENT METHODS AND EMERGING CHALLENGES

Dr. Parth Gautam[1]

[1]Associate Professor, Department of Computer Sciences and Applications, Mandsaur University, Mandsaur
parth.gautam@meu.edu.in

Abstract: The fast merging of cloud computing and artificial intelligence has essentially changed the information retrieval (IR) systems in the modern era, facilitating access of vast, intelligent and context-sensitive data. The distributed storage, parallel processing and elastic resource provisioning, which is required to handle large and heterogeneous datasets, is availed by cloud platforms, and the accuracy and efficiency of retrieval is substantially increased through algorithms in AI, including machine learning, deep learning, semantic embeddings, and neural ranking models. The following paper gives an in-depth account of cloud-based IR architecture, including the service models, distributed data storage, indexing approaches, and query processing mechanisms. It looks at the ways AI techniques and distributed learning systems can be used to maximize search at scale, and how Large Language Models (LLMs) can be applied to semantic search, query re-formulation, and retrieval-augmented generation. Also, the paper points out some of the most important ethical and technical issues, such as algorithmic bias, model drift, and privacy that influence the equitable and trustable execution of AI-based IR systems. Altogether, this paper highlights the potential of AI-assisted cloud IR as the revolution but provides the importance of responsible, secure, and scalable retrieval models in the changing data landscape.

Keywords: Cloud computing, artificial intelligence, deep learning, information retrieval, distributed architectures, privacy, model drift.

1 INTRODUCTION

The widespread adoption of cloud computing has fundamentally reshaped how data is stored, processed, and delivered across modern digital ecosystems. The current use of cloud computing has essentially transformed the ways in which data is stored, processed and delivered within contemporary digital ecosystems. Having the properties of scalability, elasticity, distributed storage, and on-demand resource provisioning, the cloud has turned out to be the foundation of modern data-driven applications [1]. With organizations ever producing huge and varied amounts of data, the necessity to derive valuable information out of the cloud setup has grown more critical, leading to the development of sophisticated retrieval systems. In this respect, the Information Retrieval (IR) is central. IR is concerned with identifying the appropriate information among collection of data which in the past relied on search using keywords, indexing schemes, and ranking functions [2]. Although classical IR methods worked well with small or more moderate-sized datasets, the dynamic and rapid increase of cloud-based data, as well as the sophisticated user search patterns requires smarter and more scalable solutions. This development has given rise to cloud-based IR, which is a paradigm where retrieval functions are executed directly inside cloud environments with the aim of exploiting distributed processing power, parallel processing and storage capacity.

Cloud-based Information Retrieval improves traditional IR as it can process large volumes of data in real time, support multi-tenant searching services, and processes heterogeneous data, document, images, logs, and multimedia [3][4]. Nevertheless, in spite of its architectural advantages, cloud-based IR continues to struggle with the intent of the user, prioritization of different content, and context-driven results, which are aggravated when data is highly complex. These gaps have necessitated the adoption of smart computing models to fill them [5][6]. This is where the Artificial Intelligence (AI) comes in. Machine and deep learning, known as AI, has strong possibilities in identifying patterns, semantics, and optimization of search context. AI is able to learn the behavior of users, to analyze the broad trends, and model non-linear relations in data, which cannot be done by traditional IR systems [7][8]. The combination of AI with cloud-based IR has led to the emergence of the AI-based Cloud Information Retrieval, a next generation in retrieval systems, which can perform semantic search and vector-based indexing, personalized ranking, and real-time relevance feedback [9]. Such deep learning methods as neural ranking models and transformer-based networks (e.g., BERT, GPT) are crucial to the richness and quality of search results. The capabilities are increased with the aid of cloud platforms which provide AI-optimized infrastructures, vector databases, as well as scalable model-serving pipelines.

1.1 Structure of the paper

This review covers cloud-based information retrieval. Section I introduces AI and cloud IR, Section II discusses cloud computing for information retrieval, Section III reviews ai techniques in cloud-based information retrieval, Section IV examines ethical challenges, Section V presents the literature review, and Section VI concludes with future directions.

2 CLOUD COMPUTING FOR INFORMATION RETRIEVAL

The Information Retrieval (IR) systems developed to be run on a distributed cloud infrastructure have a design that is set to be run on remote data, remote computation, and remote service hosting and accessed on-demand [10]. They are usually three-layered architecture, the data layer, where large-scale and heterogeneous datasets are stored, the processing layer, which processes the data with indexing, feature extraction, and searching algorithms, and the application layer, which presents user interfaces, API, and analytics tools. Distributed computing frameworks like MapReduce, Spark and containerized microservices are key to cloud IR architecture to enable parallel processing of data and fault tolerance [11][12]. Availability and resilience are guaranteed by load balancers, distributed indexes, and replication mechanisms. This scalable system allows scaling to be performed with ease whilst continuing to provide the same query performance to large datasets and across multi-region applications.

2.1 Cloud Services Innovations (IaaS, PaaS, SaaS) in IR

There are various service models that the cloud-based Information Retrieval (IR) systems can be used on such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) with unique benefits. It is shown in Figure 1.

Figure 1: Service models of cloud

- IaaS: IaaS offers virtualized compute, storage and networking capabilities, which allow organizations to develop completely customized IR pipelines [13]. Developers can use custom indexing engines, vector databases and semantic search architectures based on large-scale datasets with the availability of virtual machines, containers and distributed storage systems.
- PaaS: PaaS simplifies the development of IR since it provides controlled environments in which applications, machine learning models and search services can be deployed. Search logic can be combined with managed databases, API gateways, and serverless functions without managing infrastructure complexity.
- SaaS: SaaS is a set of pre-prepared IR solutions in the form of cloud applications. They are enterprise search engines and document discovery platforms as well as AI-based search APIs that need minimal configuration. SaaS services are particularly handy within businesses that want to acquire rapid deployment, automatic refreshing as well as low operation costs.

2.2 Data Models: Distributed File Systems and Databases

Cloud IR systems are based on powerful and elastic storage models, to manage the various types of data. Distributed File Systems include HDFS, Amazon S3 and Google Cloud storage are large stores of unstructured and semi-structured data stored in many nodes to offer high throughput and redundancy [14]. These systems allow parallel access to the data, hence quick indexing and access. Cloud-native databases are also important, besides file systems. NoSQL databases (e.g., MongoDB, Cassandra, DynamoDB) are characterized by flexibility in their schema, horizontal scalability, and low-latency access and therefore are well-suited to IR tasks such as metadata storage, session tracking, and fast key-value lookups. In the meantime, relational databases and data warehouses (e.g., Big Query, Snowflake) that are distributed are useful in providing the analytical queries and structured retrieval. Collectively, these storage models are the basis of managing the diversity, speed, and amount of information present in clouds.

2.3 Query Processing and Ranking in Cloud

The steps involved in processing queries in cloud IR systems are query interpretation, query matching with distributed indexes and ranking of the relevant documents with relevance measures [15]. Cloud environments further optimize this process by supporting the execution in parallel, caching as well as scaling indexing strategies. Indexes can be split between nodes to enable quick searching whereas sharing can facilitate the effective distribution of large datasets. Ranking models frequently also integrate classical methods (e.g., TF-IDF, BM25) such as semantic embeddings and neural ranking models. Cloud query processors are dynamic in resource allocation to meet fluctuating query demand and have feedback mechanisms, e.g., click-through data and customized preferences, to optimise ranking results. These mechanisms make sure that retrieval is fast, accurate and context oriented even in high demand settings.

2.4 Scalability, Elasticity and Cost

The key benefits of cloud IR are scalability and elasticity. Scalability enables systems to add resources including compute, storage and bandwidth, according to the volume of datasets and user traffic. Elasticity allows the upkeep or downsizing of real-time to scale, which makes resource use efficient. IR systems can use auto-scaling groups, serverless computing, and container orchestration platforms (e.g., Kubernetes) [16] to scale their capacity dynamically. But these advantages have cost implications because pricing models of the cloud are resource-based. Heavy query loads, the regular update of indexes and large amounts of storage may result in enormous costs [17]. As such, cloud IR systems have to strike a balance between the performance and cost with the help of optimization techniques like caching, tiered storage, effective indexing, and policies that scale based on workloads. These are some of the factors that determine how sustainable and economically viable cloud-based IR infrastructures should be designed.

2.5 Different Retrieval techniques

Cloud computing is a collection of computing resources used for storing or accessing data from any distant place. Fig 2 shows search schemes, Keyword search procedures are extensively used and the user is allowed to retrieve chosen data from the storage space[18].

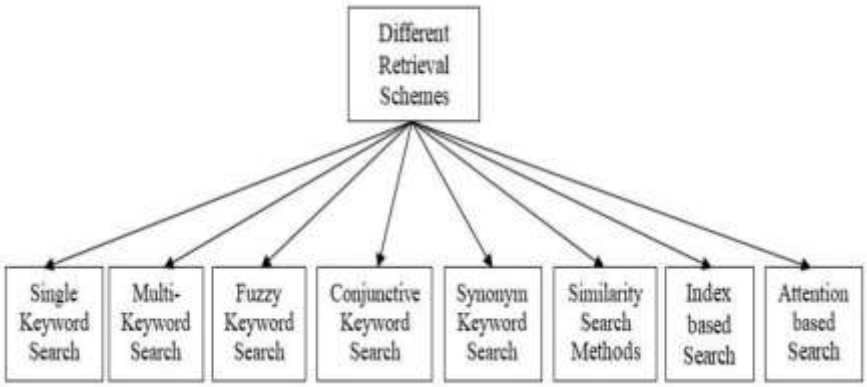

Figure 2: Different Retrieval Schemes

The various information retrieval mechanism in the cloud-based systems. Eight major approaches exist:

- **Single and multi-keyword search:** to conduct simple and combined searches.
- **Fuzzy and conjunctive search:** to retrieve similar results and complicated queries.
- **Search by synonyms:** to broaden the meaning of words.
- **Similarity-based search:** to search similar or related items.
- **Index-based search full:** speedy and effective access to data.
- **Attention-based search:** to target the most valuable information with the assistance of AI.

This classification demonstrates the extent of the ways of searching information in the cloud to ensure presence of accurate, fast, and relevant results.

3 AI TECHNIQUES IN CLOUD-BASED INFORMATION RETRIEVAL

The convergence of artificial intelligence and cloud computing has revolutionized information retrieval paradigms, enabling intelligent data access at unprecedented scales. Cloud computing, witnessing rapid innovations in recent years, has two main tasks: storing and accessing data and programs by means of Internet rather than usage of a computer's hard drive [19]. The entity cloud presents an extensive range of services. It reduces the complexity of the networks, makes provision for customization, scalability, efficiency etc. Besides, the information stored on cloud is generally not easily lost. Because of its on-demand nature, one could typically buy cloud computing the same way you would buy electricity, telephone services, or Internet access from a utility company.

3.1 Artificial Intelligence in Cloud Systems

Cloud platforms have become essential infrastructure for deploying artificial intelligence systems at scale, providing the computational resources necessary for training and inference. Machine learning (ML) and deep learning (DL) systems require scalable and effective training and inference, particularly as models continue to increase in size and complexity [20][21]. In the recent past, there has been a rising trend of implementing DNNs via cloud platforms, and cloud platform, which are high-performance computing platforms that are extremely fast and have immense memory. Cloud machine learning (ML) platforms, including Amazon Web Services (AWS) Deep Learning and Google Colab can conduct training within a reasonable time [22]. Centralized servers service involves cloud computing which offers massive computing capabilities, massive data storage, fast computation, low latency and high availability.

3.2 Distributed AI Architectures Optimizing High-Volume Data Retrieval

With the expansion of information retrieval system to work with larger datasets, the distributed AI architecture is now necessary to process and analyze large volumes of data in an efficient manner. Scalable training Scalable training is distributed training that trains

the model weights together with sharing the training data between multiple devices or nodes. On the other hand, model parallelism is used to define the division of the model into devices or nodes and the parallel computing of different (model) parts [23][24]. Distributed systems are very important in facilitating scalable and efficient machine learning (ML) processes in cloud computing. Inference in distributed systems in distributed systems, inference refers to applying a machine learning model to distributed and networked computer resources in order to obtain a prediction or a decision. This could involve conducting inference operations simultaneously with multiple nodes or devices to enhance its overall performance, latency, as well as scaling.

3.3 AI-Driven Indexing Strategies for Information Retrieval Systems

The information retrieval system or database system is majorly pegged on indexing as it is the only way of ensuring that there is a fast, accurate, and economical access to information stored. Conventional approaches of indexing such as B-trees, hash indexing and inverted indexes utilize known and fixed data structures besides present rules to sort and retrieve information [25]. However, with the increase in the size, diversity, and dynamism of data in the form of applications in big data analytics, cloud computing, and social media platforms, traditional ways of indexing are becoming inefficient. To address these deficits, AI-based indexing solutions have come up as a solution to this issue [26]. These methods can learn patterns of previous queries, correlations, and semantic relationships among items of data through the application of machine learning and deep learning algorithms.

3.4 AI Search and IR Services of Leading Cloud Providers

Getting more and more requirements in semantic search, personalization, and scalable indexing, major cloud platforms have created specialized information retrieval tools driven by AI. These services are based on machine learning, vector embeddings, and natural language understanding to improve the relevancy and efficiency of search.

- **AWS (Kendra and OpenSearch):** Amazon Kendra is an AI-based enterprise search engine that utilizes natural language understanding in finding the most accurate search results out of documents, websites, and knowledge bases [27]. It is backed by semantic ranking, question-answer, and domain-specific model tuning. Amazon OpenSearch is a cloud-managed Elasticsearch implementation, which offers full-text search, log analytics, and vector search.
- **Google Cloud (Vertex AI Search):** Vertex AI Search (previously Enterprise Search) is a Google-quality search engine that is enhanced with deep learning and NLP. It favours semantic search, vector-based retrieval, document understanding and high-quality ranking models that are trained on large-scale datasets [28]. Search services on Google Cloud can be listed as having an improved indexing algorithm, multi-linguistic support, and excellent infrastructure to support low-latency retrieval.
- **Microsoft Azure (Cognitive search):** The Azure Cognitive Search is an integration of the traditional search and the capabilities of advanced AI enrichment. It incorporates inbuilt thinking capabilities like language recognition, entity annotation, OCR and semantic ranking through deep learning frameworks. Semantic Search mode of Azure offers contextual information, query reformulation, and intent matching. It is built in with Azure Machine Learning, Form Recognizer and knowledge mining capabilities and is applicable to enterprise document search, compliance retrieval and intelligent content analytics.

3.5 Integration of LLMs in Cloud IR Pipelines

The rise of Large Language Models (LLMs) such as GPT, PaLM, and Llama has transformed cloud-based IR by enabling more intelligent and context-aware retrieval workflows. LLMs enhance IR pipelines through:

- **Semantic Embedding Generation:** LLMs are able to generate high quality dense document and query embeddings, facilitating accurate vector searches.
- **Retrieval-Augmented Generation (RAG):** Retrieval combined with LLM-based generation is effective in conditioning higher quality of answers, fewer hallucinations, and grounding of facts.
- **Query Expansion and Reformulation:** LLMs are programmed to automatically expand or refine or paraphrase queries to enhance recall and relevance.

4 ETHICAL AND TECHNICAL CHALLENGES IN AI-DRIVEN INFORMATION RETRIEVAL

The rise of AI as the main driver of modern information retrieval systems has brought critical ethical and technical challenges to the forefront, with algorithmic bias emerging as a primary concern. Information retrieval is the process of obtaining information relevant to an information need from a database. When complemented with AI, it becomes enhanced to make data interpretations and decisions [29]. An information retrieval system infused with AI can conduct proper topical searches for finding resources relevant to a query.

- **Data Privacy and Security:** AI systems frequently need user access to sensitive user information to personalize search data and improve the user experience [30]. Having privacy of the data, adherence to regulations (e.g., GDPR), and having a good level of security is a key factor.
- **Data Quality and Availability:** The quality of data used by AI-based information retrieval systems is critically important, and it should be high quality and relevant. It is important to make sure that data is accurate, complete and available in order to train AI models and produce accurate search results.
- **Scalability and Performance:** AI-based information retrieval system should be able to scale well to respond to the growing data and user requests without affecting their performance. Scalability requires optimisation of algorithms, infrastructure and allocation of resources.

- **Biases in AI algorithms:** AI algorithms can be biased depending on the datasets that they are trained, which can be skewed or discriminant. Algorithms should be neutralized and fairness in search results must be provided to various groups of users.
- **User Trust and Acceptance:** To ensure successful adoption of AI-driven systems of information retrieval, user trust and acceptance is crucial [31]. Giving clear answers to the mechanism of AI algorithms, paying attention to the preferences of users, and addressing the issues of privacy and bias is an important step towards achieving user trust.
- **Ethical Usage of AI:** The most important consideration is ethics, including transparency, accountability, and responsible AI use. Ethical practices of AI in information retrieval require setting of ethical systems, rules, and control mechanisms.
- **Interpretability and Explain ability:** AI models to be applied in information retrieval must be interpretable and explainable, particularly when making critical decisions or recommendations. Explanation of search results and how decisions are made increases transparency and trust to the users.

5 LITERATURE REVIEW

Recent studies in this section show that the use of AI in cloud information retrieval continues to grow, with its application ranging from mere accuracy and efficiency to turning in support of multimodal and domain-specific searches.

Li (2025) Cloud computing has become an important component of modern IT infrastructure and typically relies on x86 or ARM architecture hardware to provide scalable and efficient cloud services. Constructed a cloud computing cluster based on RISC-V servers and conducted extensive performance evaluations of the cluster's CPU, memory and disk. By evaluated the performance of the RISC-V cluster in containerized environments, we assess the current state and potential of RISC-V architecture in cloud computing [32].

Lin (2025) proposes a multimodal semantic enhanced attention retrieval algorithm (MSAAR). The algorithm integrates multimodal information such as text and structured data, and uses a dynamic adaptive attention mechanism to accurately retrieve the key content of literary works. The experiment selected public data sets such as classic literary masterpieces, modern literary works, and online literary works, and compared the TF-IDF algorithm and the Transformer-based BERT retrieval algorithm with accuracy, recall, and F1 value for evaluation [33].

Dhala, Kumar and Panda (2025) presents a legal document information retrieval system that retrieves the most relevant documents quickly from a collection of legal reports and documents. For this purpose, a legal document repository is created by collecting the documents and case study reports of different legal matters of last five years. The retrieval model was tested in several context to evaluate the performance of the legal model. The domain classification approach is benchmarked against a number of classifiers (K-nearest neighbours, logistic regression, random forest and XGBoost) trained in exactly the same settings [34].

Hassan et al. (2024) highlights the role of artificial intelligence in the cloud computing environment. Literature studies have identified several applications, including resource scaling, cost optimization, performance optimization, threat detection, content creation, resource allocation, and more. Different machine learning algorithms that help in cloud computing tasks have also been explored. Major machine learning algorithms for cloud computing are K-means, SVM, K-NN, Naïve Bayes, SVD, and neural networks. The major research trends have also been explored in which cloud security is the most researched theme of cloud computing [35].

Sharma et al. (2024) development of a cloud-based platform that employs artificial intelligence technology to automatically recognize and apply 11 modulation schemes (3 analogy and 8 digital) to complicated or quadrature radio signals. The SNR values under consideration are from 0.0 to 40.0 and involve moderate drift, slight fading, and labelled increments. Deep-six has forthcoming developed a large synthetic database which will train the four AI models. These will be combined with the Google Cloud AI platform to take advantage of the flexibility and processing power. The system will be tested with an SDR platform in GNU Radio, and its capabilities for real-world signal processing applications will be demonstrated. The cloud-based platforms have the enthralling features of the flexibility and computational power to supersede the traditional computers for the AI-driven signal processing [36].

Saxena, Sharma and Mehta (2024) a expansive rate of businesses depend intensely on computer innovation to realize a assortment of objectives, such as taken a toll lessening, foundation administration, improvement stages, information preparing, and information analytics. Web-based apps are advertised by cloud benefit suppliers (CSPs) to empower conclusion clients to get to administrations over the web, making it simpler for them to utilize these administrations at whatever point and wherever they need. The viability of machine learning strategies for cloud security is compared and evaluated in this investigate with respect to risk discovery and mitigation [37].

Ding and Gong (2023) the traditional information retrieval (IR) method mainly relies on the manual work of librarians, but with the development of computer technology and intelligent technology, the traditional IR method is gradually replaced by the inference mechanism of artificial intelligence (AI) by computers. Driven by big data, the scientific classification of network information can be effectively realized by applying AI technology to network IR, and the corresponding information classification framework can be set by combining users' network habits through keyword search [38].

Ren et al. (2023) cloud computing Involves various technical fields. From the perspective of the research subject, how to effectively analyze the development trend of cloud computing technology and how to formulate a cloud computing evolution route that suits Itself has become an urgent problem to be solved. Based on the classification of cloud computing products of mainstream cloud providers and consulting organizations in the Industry, Introduces a method of splitting the underlying technology of cloud products, conducting product-technology two-dimensional analysis of different technologies, and providing technical research and development suggestions from the fields of cloud Infrastructure, cloud services, cloud operations and governance, and cloud security [39].

Table 1 summarizes recent studies on AI in cloud-based information retrieval, highlighting methods, key results, challenges, and future directions, illustrating progress and ongoing issues in intelligent cloud IR systems.

TABLE I. SUMMARY OF RECENT STUDIES ON CLOUD COMPUTING & INFORMATION RETRIEVAL

Reference	Study On	Approach	Key Findings	Challenges / Limitations	Future Directions
Li (2025)	Cloud computing performance on emerging architectures	Constructed a RISC-V based cloud cluster; evaluated CPU, memory and disk performance in containerized workloads	Demonstrates that RISC-V offers competitive compute performance and energy efficiency; highlights feasibility for scalable cloud services	RISC-V ecosystem is still maturing; fewer optimizations and limited hardware support compared to x86	Optimization of RISC-V cloud stacks; broader benchmarking; integration into mainstream cloud computing environments
Lin (2025)	Multimodal information retrieval for literary works	Proposed MSAAR using text + structured data with adaptive multimodal attention	Achieves 90.2% accuracy (+12.8% vs BERT), 85.2% recall (+27.6% vs TF-IDF), and +13.5% F1 improvement across datasets	High computational overhead; requires rich multimodal datasets	Expansion to more domains; refinement of cross-lingual and multimodal retrieval
Dhala, Kumar & Panda (2025)	Legal document information retrieval	Built a legal repository; used LSTM-based domain classification; benchmarked against KNN, LR, RF, XGBoost	Reduces search space; improves retrieval efficiency and relevance across legal subdomains	Strong dependency on accurate domain labels; limited by historical case coverage	Scaling legal repositories; improving domain ontology; real-world deployment in legal research
Hassan et al. (2024)	Artificial intelligence applications in cloud computing	Literature review of ML algorithms (K-means, SVM, KNN, NB, NN, SVD) for cloud tasks	AI improves cost optimization, resource scaling, performance, and security; security remains most researched domain	Heterogeneous cloud environments; inconsistent datasets; evolving security threats	Standardizing ML-driven cloud pipelines; advanced AI-based security automation
Sharma et al. (2024)	Cloud-based AI modulation recognition	AI models trained on DeepSig synthetic SDR datasets; integrated with cloud AI platform; tested using GNU Radio	Supports 11 modulation schemes; cloud improves flexibility and compute scalability for real-time signal analysis	Performance affected by cloud latency; synthetic–real data gap; sensitivity to radio drift	Real-world SDR deployment; cloud-edge hybrid architectures for low-latency signal processing
Saxena, Sharma & Mehta (2024)	Machine-learning-based cloud security	Compared ML methods for threat detection and mitigation in cloud environments	ML improves threat prediction, anomaly detection, and risk mitigation across cloud services	Attack evolution makes static models obsolete; dataset imbalance affects accuracy	Continuous learning security models; scalable, adaptive cloud-wide threat intelligence
Ding & Gong (2023)	AI-based Internet information retrieval	Deep learning IR model enhanced by data-mining classification	Achieves 96.48% precision and 98.45% recall; significantly faster than baseline IR algorithms	Highly dependent on data quality; limited generalization to dynamic web content	More robust IR frameworks using big-data enhancement and user-behavior modelling
Ren et al. (2023)	Cloud computing technology evolution analysis	Two-dimensional product–technology mapping across cloud infrastructure, services, operations, governance and security	Provides clear cloud evolution roadmap and identifies key technological gaps	Rapid tech evolution can make mappings outdated quickly	Dynamic cloud evolution tracking frameworks; continuous technology–product alignment

6 CONCLUSION AND FUTURE WORK

The integration of cloud computing and artificial intelligence has transformed information retrieval into a highly scalable, intelligent, and context-aware process. Cloud solutions are the distributed storage, parallel processing, and scalable resource provisioning that are needed to handle the large and diverse datasets that are created in contemporary digital landscapes. These capabilities are further reinforced with the development of Large Language Models (LLMs) which have advanced semantic reasoning, query reformulation, and retrieval-augmented generation. However, despite such progress, the AI-driven IR systems continue to encounter serious issues of privacy, security, fairness, model drift, and ethical aspects of automated decision-making. These problems should be tackled, so that the cloud-based IR systems can be reliable, transparent and socially responsible in situations of high stakes and data intensive contexts.

6.1 Future Work

A further way of research should be carried out in the development of more transparent, fair and interpretable AI models that reduce algorithmic bias and enhance the reliability of retrieval results. Another important direction is privacy preserving IR architecture like federated learning, encrypted search and secure multi-party computation which facilitate the intelligent retrieval without disclosing sensitive data. Also, the growing dynamism of data environments demands self-adaptive IR systems that can cope with model drift and also adapt to changing user behavior via continuous learning. Lastly, even closer merging of LLMs with cloud IR pipelines, and in particular, with the support of vector databases, domain specific RAG models and customized semantic search, is likely to reinvent the future retrieval systems, and make it more interactive, context aware, and able to assist in real-time decision-making across multiple applications.

7 REFERENCES

[1] D. Oladimeji, K. Gupta, N. A. Kose, K. Gundogan, L. Ge, and F. Liang, “Smart Transportation: An Overview of Technologies and Applications,” *Sensors*, vol. 23, no. 8, p. 3880, Apr. 2023, doi: 10.3390/s23083880.

[2] N. F. Prangon and J. Wu, “AI and Computing Horizons: Cloud and Edge in the Modern Era,” *J. Sens. Actuator Networks*, vol. 13, no. 4, p. 44, Aug. 2024, doi: 10.3390/jsan13040044.

[3] N. Mungoli, “Scalable, Distributed AI Frameworks: Leveraging Cloud Computing for Enhanced Deep Learning Performance and Efficiency,” 2021, doi: 10.48550/arXiv.2304.13738.

[4] G. Maddali, “An Efficient Bio-Inspired Optimization Framework for Scalable Task Scheduling in Cloud Computing Environments,” *Int. J. Curr. Eng. Technol.*, vol. 15, no. 3, 2025.

[5] P. Chandrashekar and M. Kari, “Design Machine Learning-Based Zero-Trust Intrusion Identification Models for Securing Cloud Computing System,” *Int. J. Res. Anal. Rev.*, vol. 11, no. 4, pp. 901–907, 2024.

[6] S. Narang and G. K. V, “Next-Generation Cloud Security: A Review of the Constraints and Strategies in Serverless Computing,” *Int. J. Res. Anal. Rev.*, vol. 12, no. 3, 2025, doi: 10.56975/ijrar.v12i3.319048.

[7] S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir, “A Review on recent research in information retrieval,” *Procedia Comput. Sci.*, vol. 201, pp. 777–782, 2022, doi: https://doi.org/10.1016/j.procs.2022.03.106.

[8] V. Shah, “Managing Security and Privacy in Cloud Frameworks : A Risk with Compliance Perspective for Enterprises,” *Int. J. Curr. Eng. Technol.*, vol. 12, no. 6, pp. 606–618, 2022, doi: 10.14741/ijcet/v.12.6.16.

[9] R. Patel, “Advancements in Renewable Energy Utilization for Sustainable Cloud Data Centers: A Survey of Emerging Approaches,” *Int. J. Curr. Eng. Technol.*, vol. 13, no. 05, Oct. 2023, doi: 10.14741/ijcet/v.13.5.7.

[10] F. J. C. Faust *et al.*, “Embedding-based retrieval techniques for feeds,” 11960550, 2024

[11] R. Patel and P. Patel, “Machine Learning-Driven Predictive Maintenance for Early Fault Prediction and Detection in Smart Manufacturing Systems,” *ESP J. Eng. Technol. Adv.*, vol. 4, no. 1, 2024, doi: 10.56472/25832646/JETA-V4I1P120.

[12] V. Rajavel and R. Gahlot, “Advanced Fault Diagnosis of CMOS Circuit Design by Leakage Measurement in Nanometer Technology,” in *2025 IEEE 5th International Conference on VLSI Systems, Architecture, Technology and Applications (VLSI SATA)*, IEEE, May 2025, pp. 1–6. doi: 10.1109/VLSISATA65374.2025.11070065.

[13] S. Chatterjee, “A Data Governance Framework for Big Data Pipelines: Integrating Privacy, Security, and Quality in Multitenant Cloud Environments,” *Tech. Int. J. Eng. Res.*, vol. 10, no. 5, 2023, doi: 10.56975/tijer.v10i5.158181.

[14] G. Sarraf and V. Pal, “Adaptive Deep Learning for Identification of Real-Time Anomaly in Zero-Trust Cloud Networks,” vol. 4, no. 3, pp. 209–218, 2024, doi: 10.56472/25832646/JETA-V4I3P122.

[15] S. Amrale, “Proactive Resource Utilization Prediction for Scalable Cloud Systems with Machine Learning,” *Int. J. Res. Anal. Rev. (IJRAR*, vol. 10, no. 4, pp. 758–764, 2023.

[16] D. Patel, “The Role of Amazon Web Services in Modern Cloud Architecture: Key Strategies for Scalable Deployment and Integration,” *Asian J. Comput. Sci. Eng.*, vol. 9, no. 4, pp. 1–9, 2024.

[17] S. B. Karri, C. M. Penugonda, S. Karanam, M. Tajammul, S. Rayankula, and P. Vankadara, “Enhancing Cloud-Native Applications: A Comparative Study of Java-To-Go Micro Services Migration,” *Int. Trans. Electr. Eng. Comput. Sci.*, vol. 4, no. 1, pp. 1–12, Apr. 2025, doi: 10.62760/iteecs.4.1.2025.127.

[18] V. R and S. N. Chandrashekara, “A Survey on Context based Information Retrieval in Cloud,” *Int. J. Eng. Res. Technol.*, vol. 8, no. 07, pp. 153–157, 2019.

[19] A. Parupalli and H. Kali, “An In-Depth Review of Cost Optimization Tactics in Multi-Cloud Frameworks,” *Int. J. Adv. Res.*

Sci. Commun. Technol., vol. 3, no. 5, pp. 1043–1052, Jun. 2023, doi: 10.48175/IJARSCT-11937Q.

[20] K. Y. Chan *et al.*, "Deep neural networks in the cloud: Review, applications, challenges and research directions," *Neurocomputing*, vol. 545, p. 126327, Aug. 2023, doi: 10.1016/j.neucom.2023.126327.

[21] N. K. Prajapati, "Cloud-based serverless architectures: Trends, challenges and opportunities for modern applications," *World J. Adv. Eng. Technol. Sci.*, vol. 16, no. 1, pp. 427–435, 2025, doi: 10.30574/wjaets.2025.16.1.1225.

[22] B. R. Cherukuri, "Containerization in cloud computing: comparing Docker and Kubernetes for scalable web applications," *Int. J. Sci. Res. Arch.*, vol. 13, no. 1, pp. 3302–3315, Oct. 2024, doi: 10.30574/ijsra.2024.13.1.2035.

[23] S. Sadiq and S. R. M. Zeebaree, "Distributed Systems for Machine Learning in Cloud Computing: A Review of Scalable and Efficient Training and Inference," *Indones. J. Comput. Sci.*, vol. 13, no. 2, pp. 1685–1707, 2024, doi: 10.33022/ijcs.v13i2.3814.

[24] S. Thangavel, K. C. Sunkara, and S. Srinivasan, "Software-Defined Networking (SDN) in Cloud Data Centers: Optimizing Traffic Management for Hyper-Scale Infrastructure," *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.*, vol. 3, no. 1, pp. 29–42, 2022, doi: 10.63282/3050-9246.IJETCSIT-V3I3P104.

[25] V. Varma, "Secure Cloud Computing with Machine Learning and Data Analytics for Business Optimization," *ESP J. Eng. Technol. Adv.*, vol. 4, no. 3, pp. 181–188, 2024, doi: 10.56472/25832646/JETA-V4I3P119.

[26] N. Karri, S. K. Jangam, P. Sarathi, and R. Pedda, "AI-Driven Indexing Strategies," vol. 4, no. 2, pp. 111–119, 2023.

[27] B. R. Ande, "Enhancing Cloud-Native AEM Deployments Using Kubernetes and Azure DevOps," *Int. J. Commun. Networks Inf. Secur.*, vol. 15, no. 8, pp. 33–41, 2023.

[28] V. M. L. G. Nerella, "Automated Compliance Enforcement in Multi-Cloud Database Environments: A Comparative Study of Azure Purview, AWS Macie, and GCP DLP," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 4, pp. 270–283, Jul. 2025, doi: 10.32628/CSEIT25111668.

[29] M. O. Igbinovia and M. M. Danquah, "Artificial intelligence algorithm bias in information retrieval systems and its implication for library and information science professionals : A scoping review," *Tech. Serv. Q.*, vol. 42, no. 3–4, pp. 253–279, 2025, doi: 10.1080/07317131.2025.2512282.

[30] V. Shewale, "Demystifying the MITRE ATT&CK Framework: A Practical Guide to Threat Modeling," *J. Comput. Sci. Technol. Stud.*, vol. 7, no. 3, pp. 182–186, 2025, doi: 10.32996/jcsts.

[31] I. C. Martin, J. Vanschoren, and N. Polatidis, "Evolving Machine Learning : A Survey," 2025.

[32] T. Li, "Performance Evaluation of RISC-V Cloud Computing Cluster," in *2025 4th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, IEEE, Aug. 2025, pp. 285–288. doi: 10.1109/AIoTC66747.2025.11198687.

[33] M. Lin, "Research on the Efficiency of Information Retrieval of Literary Works Based on Natural Language Processing," in *2025 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, IEEE, Aug. 2025, pp. 804–809. doi: 10.1109/CIPAE66821.2025.00142.

[34] R. R. Dhala, A. V. S. P. Kumar, and S. P. Panda, "Legal Document Information Retrieval using Long Short Term Memory-Based Domain Classification Technique," in *2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3)*, 2025, pp. 1–6. doi: 10.1109/ISAC364032.2025.11156333.

[35] S. Hassan, Q. Li, M. Hassan, A. Yasin, M. Zubair, and M. Umair, "Unveiling the Application of Artificial Intelligence in Cloud Computing Environment," in *2024 5th International Conference on Innovative Computing (ICIC)*, 2024, pp. 1–6. doi: 10.1109/ICIC63915.2024.11116294.

[36] E. Sharma, R. C. Deo, C. P. Davey, B. D. Carter, and S. Salcedo-Sanz, "Poster: Cloud Computing with AI-empowered Trends in Software-Defined Radios: Challenges and Opportunities," in *2024 IEEE 25th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2024, pp. 298–300. doi: 10.1109/WoWMoM60985.2024.00054.

[37] K. Saxena, S. Sharma, and D. Mehta, "Performance Exploration of ML Algorithms in Security Facets of Cloud Computing," in *2024 1st International Conference on Sustainable Computing and Integrated Communication in Changing Landscape of AI (ICSCAI)*, 2024, pp. 1–8. doi: 10.1109/ICSCAI61790.2024.10866911.

[38] Y. Ding and R. Gong, "Design of Internet Information Retrieval System Based on Artificial Intelligence Technology," in *2023 International Conference on Telecommunications, Electronics and Informatics (ICTEI)*, 2023, pp. 264–268. doi: 10.1109/ICTEI60496.2023.00057.

[39] J. Ren, D. Fu, C. Shi, Z. Huang, W. Zhu, and Y. Liu, "Research on Cloud Computing Technology Graph Analysis," in *2023 2nd International Conference on Computing, Communication, Perception and Quantum Technology (CCPQT)*, 2023, pp. 84–91. doi: 10.1109/CCPQT60491.2023.00020.

