American Journal of Computer Science and Technology
2025, Vol. 8, No. 4, pp. 174-180
https://doi.org/10.11648/j.ajcst.20250804.11

Research Article

Evaluating Precision and Recall at Retrieval Time in Retrieval-Augmented Generation (RAG) Systems

Gopichand Agnihotram[1, *], **Joydeep Sarkar**[2]

[1]Chief Technology Office, Wipro Limited, Bengaluru, India
[2]Technology Services, Wipro Limited, Bengaluru, India

Abstract

Retrieval-Augmented Generation (RAG) systems signify a pivotal advancement in natural language processing, merging information retrieval with large language models (LLMs) to ground responses in external knowledge. This hybrid approach enhances the factual accuracy and currency of generated content, mitigating common issues like hallucination. The efficacy of a RAG system, however, is fundamentally dependent on the performance of its retrieval component. This paper provides a detailed analysis of precision and recall as critical metrics for evaluating and optimizing this retrieval step. We explore the distinct roles and inherent trade-offs of these metrics within a RAG pipeline, demonstrating their direct influence on the quality of the final output. Through a series of experiments comparing sparse (BM25), dense (DPR), and hybrid retrieval methods, we quantify their performance characteristics. The analysis is further enriched with real-world examples from finance, law, and healthcare, illustrating the practical implications of retrieval quality. Additionally, we outline advanced strategies for improving retrieval effectiveness, such as multi-stage architecture involving rerankers and the use of query transformations. The paper concludes with a set of best practices for deploying robust, enterprise-grade RAG systems, emphasizing the need for continuous evaluation and sophisticated retrieval strategies. By focusing on the systematic optimization of precision and recall, organizations can build more reliable and trustworthy AI applications.

Keywords

Retrieval-Augmented Generation (RAG), Precision and Recall, Information Retrieval, Hybrid Search, Reranking, LLM Evaluation

1. Introduction

As enterprises and research teams deploy Retrieval-Augmented Generation (RAG) systems in domains such as finance, healthcare, customer support, and compliance, there is an increasing need to evaluate and improve the retrieval components of these systems. [1, 2] RAG pipelines enable LLMs to dynamically pull in supporting documents from large, curated knowledge bases. This not only increases their coverage but also mitigates hallucination, a frequent shortcoming of purely generative models.

However, retrieval itself can introduce new challenges. Poor-quality retrieval, whether in the form of irrelevant or incomplete documents, can severely undermine the language model's ability to generate high-quality responses. Precision and recall offer a principled way to assess and refine the re-

*Corresponding author: gopichand.agnihotram@wipro.com (Gopichand Agnihotram)

Received: 12 September 2025; **Accepted:** 23 September 2025; **Published:** 18 October 2025

trieval process. High precision ensures that the context supplied to the language model is relevant and trustworthy, while high recall ensures that essential information is not omitted. Balancing these metrics is essential, yet difficult.

This paper expands upon the theory and practice of precision and recall in RAG, explaining their role, how they can be measured, and how they can be improved. We show that organizations that closely monitor and optimize these metrics can build more reliable RAG-based systems capable of supporting mission-critical tasks.

2. Understanding Precision and Recall in the Context of RAG

Precision and recall are well-established measures in the field of information retrieval. Their definitions in RAG are consistent with those in traditional search engines, but their practical impact is even more pronounced because they directly affect the language model's input context, as illustrated in Figure 1.

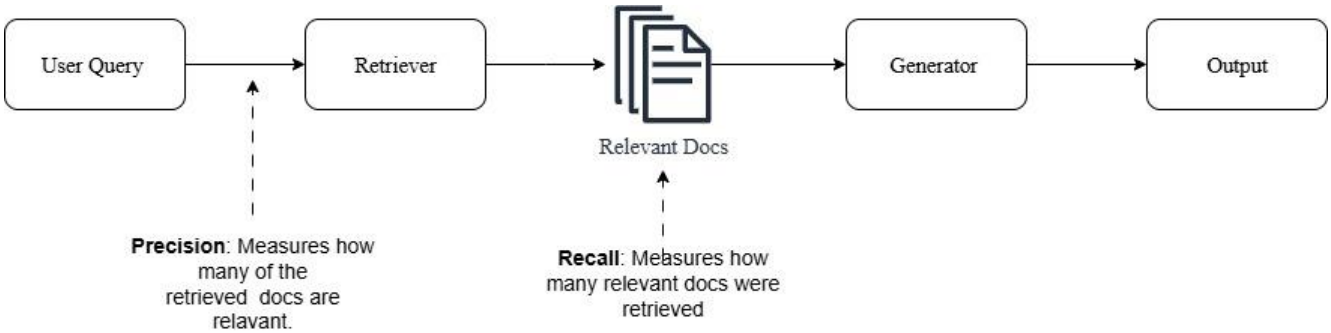

Figure 1. *Process Flow.*

Precision: The proportion of retrieved documents that are relevant to the user's query. High precision means minimal irrelevant content is introduced.

Recall: The proportion of relevant documents in the entire corpus that are successfully retrieved. High recall ensures completeness.

2.1. Precision vs Recall in RAG

These metrics form a trade-off. Retrieval systems can often be tuned to favor one at the expense of the other. For instance, retrieving a larger number of documents tends to increase recall but can dilute precision, leading to noise. Conversely, applying aggressive filters may increase precision but result in critical information being missed.

In RAG systems, the consequences of poor precision or recall are magnified. Irrelevant context wastes limited token space and confuses the LLM, increasing the chance of hallucination. Missing relevant content leads to incomplete answers and erodes user trust. Therefore, achieving the right balance is essential.

2.2. Impact on Generation Quality

The quality of the retrieval stage directly dictates the potential quality of the generation stage. If the retrieved context is flawed, even the most advanced LLM will struggle to produce a correct and complete response. Low precision introduces noise, forcing the LLM to sift through irrelevant information, which can lead to "context-stuffing" where the model incorrectly synthesizes unrelated facts. Conversely, low recall creates knowledge gaps, leaving the LLM without the necessary information to formulate a comprehensive answer, often resulting in generic or evasive responses.

3. Expanded Experimental Setup

To reflect the latest advancements in RAG, our experimental setup incorporates more sophisticated retrieval and evaluation methodologies. The landscape of retrieval has evolved, with a move towards more nuanced systems that leverage the strengths of different approaches.

BM25 (Best Matching 25)

BM25 is a classic sparse retrieval algorithm used in traditional information retrieval systems. It is built on the probabilistic relevance framework and scores documents based on the frequency and distribution of query terms within the document. The BM25 formula considers term frequency (TF), which measures how often a query term appears in a document, and inverse document frequency (IDF), which down-weights terms that are common across many documents. It also normalizes for document length, ensuring that longer documents are not unfairly favoured simply because they contain more terms. BM25 relies on an inverted index structure, making it highly efficient and scalable. However, because it depends strictly on exact keyword matching, it can struggle when the query and document use different vocabularies, leading to lower recall for semantically equivalent terms.

DPR (Dense Passage Retrieval)

Dense Passage Retrieval is a dense, embedding-based retrieval approach that leverages deep neural networks (often BERT-based encoders) to represent queries and passages as fixed-dimensional vectors in the same semantic space. During retrieval, DPR calculates the similarity between the query embedding and document embeddings, typically using cosine similarity or dot product. Unlike BM25, DPR can match on semantic similarity rather than exact keyword overlap, which makes it more robust for paraphrased queries and conceptually related terms. The retrieval pipeline involves two stages: (1) precomputing embeddings for all documents and storing them in a vector index (e.g., FAISS), and (2) encoding the query at runtime and retrieving the top-k most similar vectors. DPR offers higher recall but may reduce precision because semantically similar but less relevant documents can be retrieved.

Hybrid Retrieval

Hybrid retrieval combines the strengths of both BM25 and DPR to achieve a balance between precision and recall. [3, 11] This approach runs both retrieval pipelines, sparse (BM25) and dense (DPR), in parallel and then merges the results using rank aggregation or weighted score combination. [3] Often, a reranker (such as a cross-encoder transformer model) is applied to the merged list to reorder the results based on fine-grained relevance judgments. [4] Hybrid retrieval ensures that exact keyword matches from BM25 are not missed, while also leveraging the semantic capabilities of DPR to capture contextually relevant documents. This multi-stage pipeline generally outperforms either method alone but at the cost of increased computational complexity and latency. [3, 4]

Cost and Latency Considerations

While hybrid retrieval systems significantly enhance accuracy, they introduce important trade-offs in terms of computational cost and response time (latency). [3] Running parallel retrieval pipelines, a sparse BM25 index and a dense DPR index, inherently require more computational resources than a single-method approach. [3]

The addition of a reranking step, particularly with powerful cross-encoder models, further compounds this issue. Unlike vector search, which performs a relatively fast similarity calculation, cross-encoders process the query against each retrieved document individually through a complex neural network. This intensive computation for every query can increase retrieval latency from milliseconds to seconds, making it a critical consideration for real-time applications. [4] Therefore, organizations must balance the need for high precision and recall against the practical constraints of their budget and the user's tolerance for slower response times. [3]

To gain deeper insight into how precision and recall affect RAG, we constructed a comprehensive experimental environment spanning multiple domains and query types.

Experimental Setup Summary

Table 1. *Experimental Setup Summary.*

Component	Details
Corpus	A diverse mix of general knowledge, financial regulations, and now, multimodal data (text and images).
Query Set	Over 500 questions, including factual, analytical, and conversational queries to test a wider range of interactions.
Ground Truth	Manually annotated relevant documents for each query, forming a "golden" reference dataset for evaluation.
Retrieval Systems	BM25, DPR, Hybrid Search with a Cross-Encoder Re-ranker
Metrics	In addition to Precision@5 and Recall@5, we now include rank-aware metrics like Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (nDCG) to evaluate the order of retrieved documents. [5, 7] We also assess end-to-end quality with metrics for groundedness and contextual relevance.

Precision and Recall across Retrieval Systems

Table 2. *Precision and Recall across Retrieval Systems.*

	BM25	DPR	Hybrid
Precision	0.7	0.6	0.8
Recall	0.5	0.75	0.85

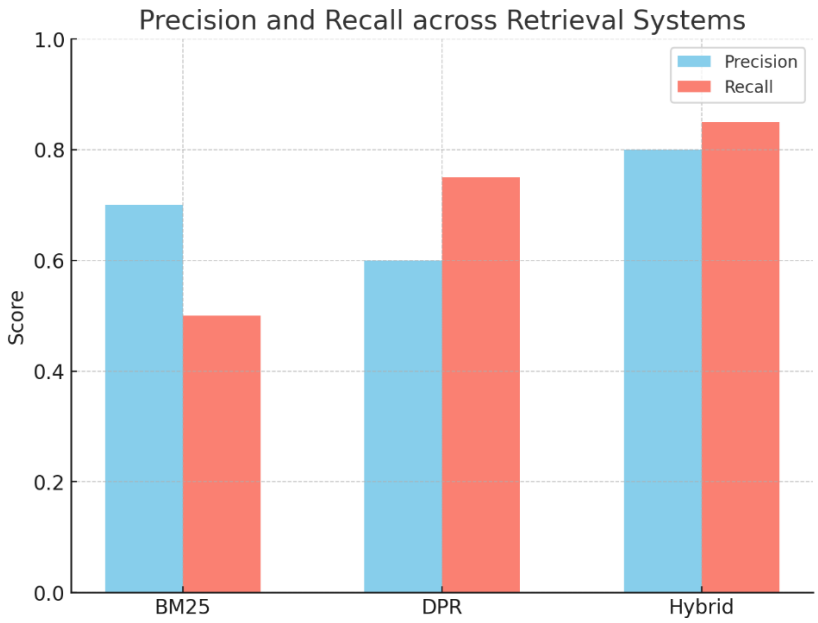

Figure 2. *Comparative view.*

Figure 2 provides a comparative visual of these results. This approximates production scenarios where RAG is used for high-stakes decision support.

4. Rich Example Walkthroughs

Example 1: SME Lending Risk Assessment

Query: "How can recent credit bureau trends be used to evaluate loan eligibility for small businesses?"

Relevant documents discuss SME lending standards, specific risk signals (e.g., late payments, credit utilization), and market-wide trends affecting small business borrowing. The DPR retriever returned a mix of relevant and tangential content. Of the top 5 retrieved documents, two directly addressed SME lending, while the rest focused on consumer credit trends and outdated financial regulations.

1) Precision@5: 40%
2) Recall@5: 60%

Example 2: EU AI Act and Autonomous Vehicle Liability

Query: "What are the implications of the EU's new AI Act for product liability in autonomous vehicles?"

Using a hybrid retriever with reranking, the system surfaced four highly relevant documents covering AI governance, liability clauses for vehicle manufacturers, and enforcement mechanisms. One retrieved document was somewhat general, discussing EU digital regulations.

1) Precision@5: 80%
2) Recall@5: 90%

Example 3: Clinical Trials and Drug Safety (Healthcare Domain)

Query: "What recent clinical trial data supports the safety of mRNA-based cancer vaccines?"

Here, the hybrid system retrieved a broad array of clinical trial results. Two documents provided updated safety findings, while others referenced preliminary studies from unrelated therapeutic areas.

1) Precision@5: 60%
2) Recall@5: 85%

Precision and Recall of Example Queries

Table 3. *Precision and Recall of Example Queries.*

Query	Precision@5	Recall@5
SME Lending Risk Assessment	40%	60%
EU AI Act and Autonomous Vehicle Liability	80%	90%
Clinical Trials and Drug Safety	60%	85%

5. Expanded Analysis of Precision and Recall Across Systems

Our updated analysis reveals a more nuanced picture of the trade-offs between precision and recall in modern RAG systems. While the fundamental tension remains, newer techniques offer more sophisticated ways to manage it.

The inclusion of a reranking stage after initial retrieval has proven to be highly effective at boosting precision without a significant drop in recall. [4] The first stage is optimized for high recall, ensuring all potentially relevant documents are captured. The re-ranker then filters and reorders this set, promoting the most relevant documents to the top. This two-stage process allows for a better balance than single-stage retrieval methods. [6]

Precision vs. Recall Trade-off with Advanced Systems

A visual representation of our findings (Figure 3) would show DPR with moderate precision and high recall, while a simple hybrid search improves precision. The standout performer is the two-stage retrieval with a re-ranker, which pushes the system closer to the ideal top-right quadrant, indicating both high precision and high recall. Adaptive retrieval systems show variable performance, intelligently shifting between high precision for simple queries and high recall for more complex ones, optimizing resource usage.

Furthermore, we've observed that the definition of "relevance" itself is becoming more sophisticated. It's not just about semantic similarity; it's about the contextual utility of the retrieved information for the specific generation task. This highlights the need for evaluation metrics that go beyond simple precision and recall measuring the actual contribution of retrieved documents to the final generated output. [5, 7]

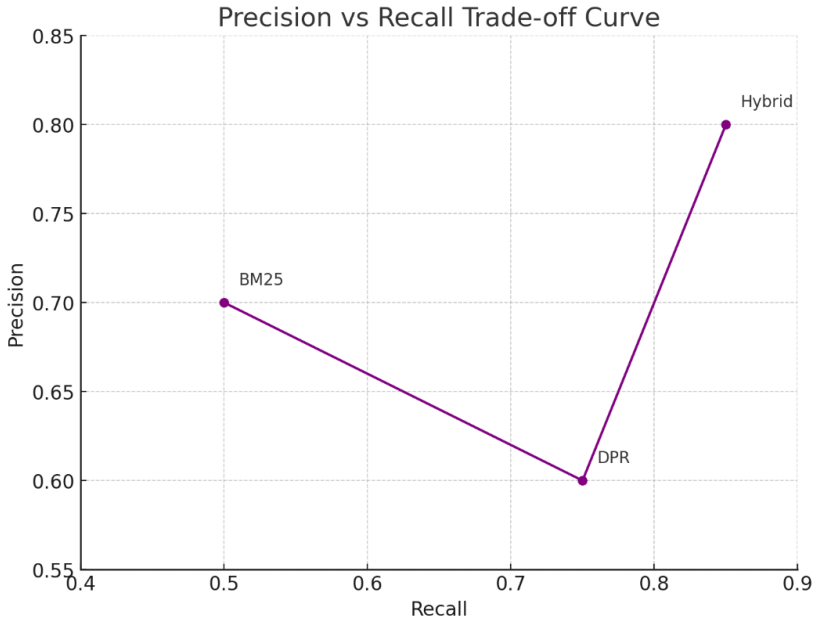

Figure 3. *Performance plot.*

6. Enhancing Retrieval Effectiveness Through Precision and Recall

Precision and recall should not only serve as evaluation metrics but also guide improvements in the retrieval pipeline. A multi-stage process, as shown in Figure 4, is a common strategy where initial retrieval prioritizes high recall, and a subsequent reranking step improves precision. [6]

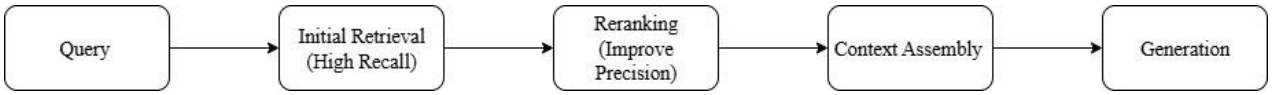

Figure 4. *Improvement plan.*

Strategies:

1) Dynamic Threshold Tuning
2) Precision-Aware Reranking [4]
3) Query Expansion & Transformation [8, 9]

4) Multi-Stage Retrieval [6]
5) Feedback Loops
6) Domain-Specific Metadata
7) Reinforcement Learning

7. Best Practices for Enterprise RAG Systems

For enterprises deploying RAG systems in production, the focus has shifted from basic implementation to continuous and rigorous evaluation to prevent "silent failures" that can erode user trust. Here are the updated best practices:

1) Establish a Robust Evaluation Framework: Create a comprehensive and automated testing framework before deployment. This should include a high-quality, diverse test set of questions and a "golden" set of desired outputs for benchmarking. [5, 10]
2) Embrace Hybrid and Multi-Stage Retrieval: Rely on hybrid search (combining keyword and vector search) as a baseline. [3, 11] For higher accuracy, implement a two-stage retrieval process with a powerful re-ranker to optimize for both high recall in the initial stage and high precision in the final context provided to the LLM. [4, 6]
3) Continuously Monitor and Iterate: RAG systems are not static. Implement a continuous refresh pipeline for your knowledge base to keep information current. Regularly monitor key metrics like retrieval accuracy, response groundedness, and latency through dashboards to quickly identify and address issues. [5, 7]
4) Incorporate Advanced Query Techniques: Improve retrieval by using query expansion and transformation techniques. [8, 9] This involves methods like generating hypothetical document embeddings or breaking down complex questions into sub-queries to retrieve more targeted results.
5) Leverage User Feedback and Domain Adaptation: Implement feedback loops to capture user interactions, which can be invaluable for fine-tuning retrieval and reranking models. [12] Fine-tune embedding models on your specific domain data to improve the understanding of niche terminology and concepts.
6) Explore Knowledge Graphs and Multimodality: For complex information landscapes, consider integrating knowledge graphs to capture relationships between entities, enabling more sophisticated retrieval strategies. [13] As RAG systems evolve, the ability to retrieve and synthesize information from multiple modalities, such as text and images, is becoming increasingly important. [14, 15]

8. Conclusion

Precision and recall form the backbone of retrieval quality in RAG systems. When carefully measured and optimized, they enable language models to operate with the right balance of relevance and completeness, directly improving factual accuracy and user trust. Future research should explore adaptive retrieval strategies and context-aware retriever fine-tuning to dynamically balance these metrics.

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
FAISS	Facebook AI Similarity Search

Author Contributions

Gopichand Agnihotram: Conceptualization, Resources, Validation, Writing - original draft, Writing - review & editing

Joydeep Sarkar: Conceptualization, Resources, Validation, Writing - original draft, Writing - review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Gao, Y., et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv: 2312.10997*. https://doi.org/10.48550/arXiv.2312.10997

[2] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33*.

[3] Hofstätter, S., et al. (2023). Fid-kd: A knowledge distillation approach for efficient and effective hybrid retrieval. *arXiv preprint arXiv: 2305.07443*.

[4] Ma, Y., et al. (2023). "Rerank is all you need" for retrieval-augmented generation. *arXiv preprint arXiv: 2311.01633*.

[5] Saad-Falcon, J., et al. (2024). ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 338-354). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.20

[6] Asai, A., et al. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

[7] Es, S., et al. (2023). Ragas: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint arXiv: 2309.15217*. https://doi.org/10.48550/arXiv.2309.15217

[8] Soni, A., & Potti, N. (2023). Query Expansion for Retrieval-Augmented Generation. *Medium*.

[9] Shamasna, S. (2024). RAG II: Query Transformations. *The Deep Hub. Medium*.

[10] Chen, J., et al. (2024). Benchmarking Large Language Models in Retrieval-Augmented Generation. *arXiv preprint*.

[11] Zhuang, S., et al. (2024). A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning. *arXiv preprint*.

[12] Jiang, Z., et al. (2023). Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.495

[13] Wang, H., & Shi, Y. (2024). Knowledge Graph Combined with Retrieval-Augmented Generation for Enhancing LMs Reasoning: A Survey. *Academic Journal of Science and Technology*. https://doi.org/10.54097/h21fky45

[14] Thiyagarajan, K. (2025). Multimodal RAG for Enhanced Information Retrieval and Generation in Retail. In *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*. https://doi.org/10.1109/ICVADV63329.2025.10961713

[15] O'Brien, C. (2024). Multimodal RAG: Beyond Text-Based Search. *Medium*.

Biography

Gopichand Agnihotram is a seasoned AI and ML leader with more than 18 years of experience in research, innovation, and enterprise-scale solution delivery. His career spans leading IT services and product organizations, where he has been instrumental in shaping AI strategies, building high-impact platforms, and driving digital transformation initiatives across diverse industries such as healthcare, BFSI, retail, automotive, and industrial systems. He has made significant intellectual contributions as the holder of 34 patents (23 granted U.S. patents) and as the author of over 36 international research publications and book chapters in reputed forums including IEEE, ACM, IET, and Springer. His scholarly work and applied innovations bridge advanced statistical modeling, reliability engineering, and AI-driven automation, reflecting a rare combination of academic rigor and industry relevance. Beyond research, Dr. Agnihotram has demonstrated strong leadership in guiding AI divisions, mentoring interdisciplinary teams, and aligning innovation with customer and organizational success. He has successfully led practices that delivered large-scale AI solutions deployed across hundreds of enterprise accounts worldwide, while also nurturing the next generation of researchers and practitioners by mentoring postgraduate students, interns, and early-career professionals. Widely recognized for his contributions, he has received multiple awards for innovation, delivery excellence, and thought leadership. His professional activities extend to serving as a reviewer for leading journals, contributing to conference program committees, and guiding academic research. With a Ph.D. in Statistics from IIT Bombay and advanced training in digital transformation from IIM Indore, he brings a unique blend of technical depth, strategic vision, and execution excellence in the evolving AI landscape.

Joydeep Sarkar is a Distinguished Member of Technical Staff (Principal Member) and Senior Architect at Wipro Limited. With over twenty years of extensive experience in artificial intelligence, machine learning, and distributed computing, he has consistently contributed to cutting-edge technological advancements. Joydeep is proficient in multiple programming languages, enabling him to design and implement robust, scalable solutions. His dedication to innovation is further demonstrated through the filing of several patents in the domains of distributed computing and AI/ML. Additionally, he has authored numerous articles published in prestigious journals, underscoring his expertise and commitment to advancing the field. Joydeep has a keen interest in security and deepfake detection, areas where he actively pursues research and development. He is a valued member of the Responsible AI Taskforce, contributing to ethical and responsible deployment of AI technologies. His current research spans Artificial Intelligence, Generative AI, Agentic AI, and fake voice detection. Committed to driving innovation and mentoring emerging talent, Joydeep continues to play a pivotal role in shaping the future of AI and distributed systems.

Research Field

Gopichand Agnihotram: Artificial Intelligence, Generative AI, Agentic AI, Fake voice detection
Joydeep Sarkar: Artificial Intelligence, Generative AI, Agentic AI, Fake voice detection