International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

Knowledge-Adaptive Question Answering with Retrieval-Augmented Generation on Amazon Bedrock

Duru Juliet Chinenye*, Ogbuagu Chinedu Samuel**, Chima Aguocha Obingonye***, Hilary Prince-Daniel Chikaodi****

*(Department of ICT, Abia State University, Uturu, Nigeria
Email: duru.juliet@abiastateuniversity.edu.ng)

**(Department of Computer Engineering, Enugu State University of Science and Technology (ESUT), Nigeria
Email: ogbuaguchinedusamuel@gmail.com)

***(Department of Computer Engineering, University of Derby, United Kingdom
Email: obingonyechima@gmail.com)

****(Department of Computer Engineering, Abia State University,Uturu
Email: Hilarydan98@gmail.com)

Abstract:

Knowledge-intensive question answering in high-stakes domains such as medicine, law, and finance demands systems that deliver accurate, verifiable, and temporally current information. While large language models (LLMs) exhibit remarkable generative fluency, they are fundamentally constrained by fixed knowledge cutoffs, hallucination, and limited source attribution — rendering them unreliable for decision-critical applications where factual precision, regulatory compliance, and auditability are essential. This paper presents a Knowledge-Adaptive Retrieval-Augmented Generation (KA-RAG) framework implemented on Amazon Bedrock, addressing these limitations by grounding generative responses in dynamically maintained, domain-partitioned retrieval indices rather than static parametric memory. The framework integrates four core components: hybrid retrieval combining Titan Embeddings dense search with OpenSearch BM25; cross-encoder re-ranking for precision-optimised passage selection; dynamic knowledge fusion for multi-document context assembly; and citation-augmented generation with Claude 3 and Llama 3, producing responses with token-level source attribution for full verifiability.

A central design principle of KA-RAG is knowledge adaptivity: incoming queries are automatically classified by domain and routed to independently maintained, versioned knowledge partitions, ensuring retrieval operates over domain-coherent corpora rather than mixed-domain indices. This design respects domain-specific data governance and multi-tenant isolation requirements critical for regulated industries. Bedrock Guardrails are integrated as a factuality verification layer, applying multi-stage consistency checks between retrieved evidence and generated claims to suppress hallucinated content before responses reach end users. The serverless Bedrock infrastructure further eliminates provisioning overhead and reduces knowledge update latency from days to hours through automated ingestion pipelines.

Experiments across three benchmark datasets — MedQA (medical), LegalBench (legal), and FinQA (financial) — demonstrate substantial improvements over competitive RAG baselines: a 34.2% increase in exact match accuracy, a 28.7% improvement in factual consistency, and a 41.3% reduction in hallucination rates. Ablation studies confirm that each architectural component contributes independently to overall gains. The framework's token-level attribution mechanism additionally enables compliance officers to audit every response against its source evidence — a capability increasingly demanded by regulatory frameworks governing AI in healthcare and financial services. These results establish that cloud-native, retrieval-augmented architectures designed with domain adaptivity and operational transparency as first-class concerns can meet the reliability standards required for enterprise deployment.

Keywords — **Retrieval-Augmented Generation, Question Answering, Amazon Bedrock, Knowledge Bases, Hallucination Mitigation, Domain Adaptation, Large Language Models**

I. INTRODUCTION

A. Motivation

Large language models (LLMs) have transformed the landscape of question answering, yet they remain constrained by structural limitations: knowledge cutoff, hallucination, domain specificity, verifiability, and scalability. These limitations become especially pronounced in high-stakes domains such as medicine, law, and finance, where factual precision and source traceability are essential.

Amazon Bedrock provides a cloud-native environment that directly addresses these challenges through continuously updated Knowledge Bases with automated ingestion pipelines, model-agnostic RAG orchestration using Claude, Llama, Titan, and Mistral, Guardrails for safety, compliance, and hallucination mitigation, scalable vector search via Titan Embeddings and OpenSearch Serverless, and enterprise-grade security and auditability.

These capabilities position Amazon Bedrock as an ideal substrate for knowledge-adaptive question answering — enabling systems that are not only accurate and up to date but also transparent, verifiable, and operationally scalable. This alignment between architectural needs and platform capabilities forms the foundation for the KA-RAG framework proposed in this work.

II. RELATED WORK

A. Question Answering Systems

Research in QA has evolved through several paradigms. Early systems were rule-based, relying on handcrafted linguistic patterns and domain-specific ontologies. The emergence of deep learning introduced three dominant QA families: Extractive QA (BERT, RoBERTa, ALBERT), Open-domain QA (DrQA, DPR), and Generative QA (T5, BART, GPT-4). This progression highlights a shift from deterministic rule-based systems to probabilistic, retrieval-augmented, and generative architectures capable of handling complex, multi-document reasoning.

B. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) emerged as a response to the limitations of purely parametric LLMs. Foundational work includes REALM (end-to-end differentiable retrieval), RAG by Lewis et al. (hybrid DPR + BART), FiD (Fusion-in-Decoder) for multi-document attention, and RETRO (chunk-level retrieval with cross-attention). Amazon Bedrock extends these ideas with multi-model RAG pipelines, serverless vector search, and automatic document ingestion through Bedrock Knowledge Bases.

C. Dense Retrieval Methods

Dense retrieval has become a cornerstone of modern QA. Key methods include DPR (dual-encoder with contrastive learning), ColBERT (late-interaction token-level matching), ANCE (approximate nearest-neighbor negative sampling), and SimCSE (contrastive sentence embeddings). In a Bedrock-native context, these are complemented by Titan Embeddings G1 and vector quantization with ANN search in OpenSearch Serverless.

D. Domain-Specific NLP

Domain adaptation is essential for high-stakes QA tasks. Amazon Bedrock enhances domain adaptation through Custom Model fine-tuning, domain-specific embedding spaces, and multi-tenant knowledge isolation — ensuring sensitive medical, legal, or financial data remains segregated and compliant.

E. Hallucination Mitigation

Hallucination remains one of the most critical challenges in generative QA. Bedrock introduces Guardrails for safety and factuality constraints, model invocation tracing for token-level attribution, and confidence-aware generation, allowing systems to abstain when uncertainty is high. These features significantly reduce unsupported claims and improve trustworthiness.

III. METHODOLOGY

A. System Architecture on Amazon Bedrock

The system follows a six-stage cloud-native pipeline: Query → Retrieval → Re-ranking → Generation → post-processing → Answer. This design leverages Bedrock's managed foundation models, Knowledge Bases, vector search, and orchestration capabilities to create a scalable, maintainable, and verifiable RAG system.

1) Query Processing: A Bedrock Agent receives the user query and performs semantic expansion using Titan Text or Claude. 2) Hybrid Retrieval: Dense retrieval via Titan Embeddings G1 in OpenSearch Serverless is combined with BM25 sparse retrieval using Reciprocal Rank Fusion (RRF). 3) Re-ranking: A cross-encoder fine-tuned on Bedrock evaluates each query-document pair jointly. 4) Dynamic Knowledge Fusion: Retrieved documents are combined using attention-based weighting. 5) Citation-Augmented Generation: Claude 3 or Llama 3 generates the final answer with token-level attribution. 6) Post-processing: NLI-based factuality checks, hallucination detection, and confidence scoring.

B. Hybrid Retrieval Module

The hybrid retrieval module integrates dense and sparse retrieval. Dense retrieval uses Titan Embeddings to compute cosine similarity. Sparse retrieval uses BM25 based on term frequency and IDF. Reciprocal Rank Fusion combines rankings: $RRF(d) = \Sigma\ 1/(k + r(d))$. This hybrid approach mitigates weaknesses of each retrieval type across domains with varied terminology.

C. Context-Aware Re-ranking

The cross-encoder re-ranker incorporates multi-task ranking objectives combining pointwise, pairwise, and listwise losses. Latency-accuracy tradeoffs are addressed by limiting re-ranking to the top-k retrieved documents. Bedrock

fine-tuning workflows enable secure, domain-specific adaptation without exposing sensitive data outside AWS.

D. Dynamic Knowledge Fusion

Each document receives attention weight $\alpha_i = \exp(s_i/\tau) / \Sigma\exp(s_j/\tau)$, where s_i is the re-ranking score and τ controls distribution sharpness. Multi-head attention allows the generator to attend to different semantic aspects simultaneously. Fusion stability analysis ensures robustness across noisy retrieval results.

E. Citation-Augmented Generation

Token-level attribution maps each generated token to the document segment with the highest attention weight, enabling inline citations [1-3]. Faithfulness loss during fine-tuning penalizes claims not supported by retrieved evidence. A NLI verification layer provides a final factuality score.

F. Post-processing and Verification

Ensemble NLI models evaluate entailment, contradiction, and neutrality. Contradiction detection pipelines flag inconsistencies between the generated answer and retrieved evidence. Confidence calibration curves adjust model confidence scores to better reflect actual reliability. A hallucination taxonomy categorizes errors into retrieval failures, reasoning errors, unsupported claims, and synthesis issues.

IV. EXPERIMENTAL SETUP

A. Datasets

The evaluation spans three high-stakes domains. Medical Domain: MedQA contains medical licensing exam questions requiring clinical reasoning and evidence-based retrieval from PubMed abstracts and clinical guidelines. Legal Domain: LegalBench consists of case-based reasoning tasks, statutory interpretation, and precedent-driven queries from court opinions and legal codes. Financial Domain: FinQA focuses on numerical reasoning, financial statement interpretation, and temporal analysis from SEC filings and earnings reports.

B. Baseline Systems

Baselines include: BM25 + Extractive QA (traditional IR with BERT extraction), DPR + Extractive QA (dense retrieval with extractive reading), GPT-4 Zero-shot (pure parametric generation), GPT-4 + Simple RAG (basic retrieval-augmented prompting), Atlas (state-of-the-art retrieval-augmented model), and Self-RAG (self-reflective RAG with retrieval decisions).

C. Bedrock Infrastructure

The system is deployed entirely on Amazon Bedrock. Query Encoder: Titan Embeddings G1. Document Store: OpenSearch Serverless with hybrid BM25 + ANN indexing. Re-ranker: Cross-encoder fine-tuned via Bedrock Custom Model training. Generator: Claude 3 Sonnet or Llama 3 70B. NLI Model: Bedrock-hosted DeBERTa-style model. Retrieval configuration: top-100 for hybrid search, top 20 after re-ranking, top-5 for generation context with HNSW-based ANN search.

D. Evaluation Metrics

Answer Quality: Exact Match (EM), F1 Score, ROUGE-L, BERTScore. Factual Consistency: NLI Entailment, Citation Accuracy, Hallucination Rate, Source Overlap. Retrieval Performance: Recall@k, MRR, NDCG@k, Precision@k. Efficiency: Latency, Throughput, Index Size, Computational Cost.

V. RESULTS AND ANALYSIS

A. Overall Performance Across Domains

The Bedrock-native KA-RAG system demonstrates consistent improvements across all evaluated domains. Results in Table I show a 34.2% improvement in Exact Match over traditional IR-based baselines, an 11.7% improvement over Self-RAG (the strongest prior baseline), and substantial gains in F1, ROUGE-L, and BERTScore across all domains.

TABLE I: OVERALL PERFORMANCE COMPARISON ACROSS DOMAINS

System	Med EM	Med F1	Legal EM	Legal F1	Fin EM	Fin F1
BM25 + Extractive QA	32.4	41.2	28.9	37.5	24.6	33.8
DPR + Extractive QA	38.7	48.3	34.2	43.6	29.3	39.1
GPT-4 (Zero-shot)	45.2	54.8	41.7	51.2	38.4	47.9
GPT-4 + Simple RAG	52.8	62.1	48.3	57.6	44.7	54.2
Atlas	58.3	67.4	53.6	62.9	49.8	59.3
Self-RAG	61.7	70.2	56.4	65.7	52.3	61.8
KA-RAG (Bedrock)	**68.9**	**76.8**	**63.7**	**72.4**	**58.6**	**68.1**

B. Factual Consistency and Hallucination Reduction

The system integrates multiple verification layers — NLI checks, contradiction detection, and token-level attribution — to ensure generated answers remain grounded in retrieved evidence. As shown in Table II, NLI entailment reaches 89.7%, citation accuracy 86.4%, and hallucination rate drops to 8.9% — a 41.3% reduction versus the best baseline.

TABLE II: FACTUAL CONSISTENCY AND HALLUCINATION METRICS

System	NLI Entail (%)	Citation Acc (%)	Halluc Rate (%)	Source Overlap (%)
GPT-4 (Zero-shot)	62.3	—	31.4	—
GPT-4 + Simple RAG	71.8	68.2	24.7	73.5
Atlas	78.4	74.6	18.9	79.2
Self-RAG	82.1	78.3	15.2	82.7

KA-RAG (Bedrock)	89.7	86.4	8.9	91.3

C. Retrieval Performance

The hybrid retrieval module achieves high recall and ranking precision across domains. As shown in Table III, dense retrieval captures semantic similarity, sparse retrieval captures rare terms and identifiers, and RRF balances both signals. Cross-encoder re-ranking provides the largest single-stage improvement in precision.

TABLE III : RETRIEVAL PERFORMANCE ACROSS METHODS

Method	Recall@5	Recall@20	Recall@100	MRR	NDCG@20
BM25	42.3	61.7	78.4	0.523	0.612
Dense (DPR)	51.8	72.4	86.2	0.614	0.698
Hybrid (Dense+ Sparse)	58.7	79.3	91.6	0.682	0.751
Hybrid+ Cross-Encoder	64.2	84.1	91.8	0.738	0.809

D. Ablation Studies

Ablation experiments in Table IV isolate each component's contribution. Removing re-ranking produces the largest drop in EM and F1. Removing dynamic fusion increases hallucination rates due to noisy context aggregation. Removing citation mechanisms reduces trustworthiness. Table V presents LegalBench domain-specific ablations.

TABLE IV: ABLATION STUDY RESULTS (MEDICAL DOMAIN)

Configuration	EM	F1	NLI Entail (%)	Halluc Rate (%)
Full KA-RAG System	68.9	76.8	89.7	8.9
Hybrid Retrieval	64.2	72.1	87.3	11.2
Re-ranking	62.8	70.4	85.6	13.7
Dynamic Fusion	65.3	73.5	86.9	12.4
Citation Attribution	67.4	75.2	88.1	10.6
Faithfulness Loss	66.1	74.3	84.2	14.8

TABLE V: ABLATION STUDY ON LEGALBENCH

Configuration	Statute Interp	Case Reasoning	Precedent Match	Overall
Full System	71.4	68.2	74.1	71.2
Re-ranking	65.7	61.3	68.4	65.1
Dynamic Fusion	67.2	63.8	70.1	67.0
Citation Attribution	69.1	65.4	72.3	68.9

E. Domain-Specific Behavior

Medical QA benefits most from semantic retrieval and citation grounding for guideline-based questions. Legal QA shows strong improvements in precedent retrieval and statute interpretation, with citation-aware generation improving legal traceability. Financial QA gains from accurate numerical extraction and temporal reasoning with structured document retrieval.

F. Efficiency and Scalability

Table VI presents latency breakdown. Query processing and retrieval remain fast (18ms and 42ms) due to serverless vector search. Re-ranking adds 71ms but significantly improves accuracy. Generation with Claude 3/Llama 3 is the dominant cost at 78ms. Total end-to-end latency averages 300ms. OpenSearch Serverless scales automatically; Bedrock Agents orchestrate multi-model pipelines without manual provisioning.

TABLE VI : LATENCY BREAKDOWN ACROSS PIPELINE STAGES (MS)

Pipeline Stage	Mean Latency (ms)	Std Dev	Contribution (%)
Query Expansion	18	4	6.2
Dense Retrieval (Titan)	42	7	14.5
Sparse Retrieval (BM25)	36	6	12.4
Reciprocal Rank Fusion	9	2	3.1
Cross-Encoder Re-ranking	71	11	24.5
Dynamic Knowledge Fusion	28	5	9.7
Generation (Claude/Llama)	78	13	26.9
Post-processing (NLI)	18	3	6.2
Total	300 ms	—	100%

G. Error Analysis

A detailed error taxonomy in Table VII reveals four major categories. Retrieval failures account for 27.4% of errors, reasoning errors for 22.1%, unsupported claims for 18.3%, and citation misalignment for 14.8%. Ambiguous queries (9.6%) and formatting errors (7.8%) comprise the remainder.

TABLE VII: ERROR TAXONOMY DISTRIBUTION

Error Type	Description	Frequency (%)
Retrieval Failure	Relevant document not retrieved	27.4
Reasoning Error	Multi-hop or numerical reasoning failure	22.1
Unsupported Claim	Generated content not grounded in evidence	18.3
Citation Misalignment	Incorrect or missing citation	14.8
Ambiguous Query	Misinterpretation of user intent	9.6
Formatting Error	Missing sections, incomplete answer	7.8

H. Human Evaluation

Domain experts in medicine, law, and finance evaluated the system on accuracy, completeness, relevance, citation quality, and trustworthiness. Experts consistently rated KA-RAG higher than generic LLMs. Citation quality was particularly appreciated in legal and medical domains. Financial analysts valued accurate numerical extraction.

VI. DISCUSSION

A. Architectural Insights

Hybrid retrieval improves robustness across heterogeneous domains by combining dense semantic signals with sparse lexical matching. Cross-encoder re-ranking provides the largest single-stage accuracy gain by capturing fine-grained query-document interactions. Dynamic knowledge fusion with attention-based weighting reduces noise and enhances factual grounding. Citation-augmented generation increases transparency and trust through token-level attribution, essential in regulated domains.

B. Limitations

Cross-encoder re-ranking and large-model generation introduce latency and cost. Retrieval performance is constrained by knowledge base completeness and freshness. The system struggles with multi-hop causal chains, legal precedent synthesis, and multi-variable financial reasoning. High-quality domain adaptation requires labeled data and domain expertise.

C. Ethical and Safety Considerations

Medical QA systems must be positioned as informational tools, not diagnostic engines. Legal QA must avoid providing definitive legal advice given jurisdictional differences. Financial QA must avoid speculative interpretations. Retrieved documents may contain systemic biases that propagate into generated answers, requiring mitigation. Bedrock's multi-tenant isolation and encryption help protect sensitive domain data.

D. Future Directions

Future work should explore improved multi-hop reasoning via chain-of-thought prompting and graph-based reasoning, real-time knowledge updates with incremental indexing and temporal reasoning, uncertainty quantification through Bayesian inference and ensemble scoring, multilingual and cross-lingual QA with cross-lingual embeddings, and conversational multi-turn QA with context retention.

VII. CONCLUSIONS

The Knowledge-Adaptive Retrieval-Augmented Generation (KA-RAG) framework implemented on Amazon Bedrock demonstrates that domain-specific question answering can be both highly accurate and operationally scalable when retrieval, re-ranking, dynamic fusion, and citation-aware generation are tightly integrated within a managed cloud ecosystem. The system consistently outperforms traditional IR-based QA, dense-only retrieval, and prior RAG architectures across medical, legal, and financial domains, achieving a 34.2% improvement in exact match accuracy, 28.7% improvement in factual consistency, and 41.3% reduction in hallucination rates.

The Bedrock-native implementation provides serverless vector search, automated knowledge ingestion, multi-model orchestration, and enterprise-grade governance, reducing operational burden and enabling continuous knowledge updates without retraining. Despite challenges in multi-hop reasoning and domain adaptation costs, this work demonstrates that combining RAG with Amazon Bedrock's managed AI ecosystem provides a powerful foundation for trustworthy, domain-adaptive QA systems suitable for knowledge-intensive and safety-critical domains.

ACKNOWLEDGMENT

The authors acknowledge the contributions of the open-source research community in developing the datasets and baseline systems used in this evaluation. The authors also thank Amazon Web Services for providing access to the Bedrock platform for experimental evaluation.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019.

[2] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," *Proceedings of ACL*, 2017.

[3] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *NeurIPS*, 2020.

[4] T. Brown, B. Mann, N. Ryder, et al., "Language Models Are Few-Shot Learners," *NeurIPS*, 2020.

[5] G. Izacard and E. Grave, "Leveraging Passage Retrieval With Generative Models for Open-Domain Question Answering," *ACL*, 2021.

[6] S. Borgeaud, A. Mensch, J. Hoffmann, et al., "Improving Language Models by Retrieving From Trillions of Tokens," *ICML*, 2022.

[7] V. Karpukhin, B. Oguz, S. Min, et al., "Dense Passage Retrieval for Open-Domain Question Answering," *EMNLP*, 2020.

[8] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction," *SIGIR*, 2020.

[9] L. Xiong, C. Xiong, Y. Li, et al., "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval," *ICLR*, 2021.

[10] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," *EMNLP*, 2021.

[11] S. Gururangan, A. Marasovic, S. Swayamdipta, et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *ACL*, 2020.

[12] J. Howard and S. Ruder, "Universal Language Model Fine-Tuning for Text Classification," *ACL*, 2018.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT: A Distilled Version of BERT," *NeurIPS Workshop*, 2019.

[14] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the Factual Consistency of Abstractive Summarization," *EMNLP*, 2020.

[15] K. Shuster, S. Poff, M. Chen, et al., "Retrieval-Augmented Generation for Knowledge-Grounded Dialogue," *EMNLP*, 2021.

[16] J. Kuhn, et al., "Uncertainty-Aware Language Models," *arXiv:2302.09664*, 2023.

[17] H. Rashkin, et al., "Measuring Attribution in Language Models," *ACL*, 2023.

[18] A. Asai, et al., "Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection," *arXiv:2310.11511*, 2023.

[19] G. Izacard, et al., "Atlas: Few-Shot Learning With Retrieval-Augmented Language Models," *arXiv:2208.03299*, 2022.

[20] K. Guu, K. Lee, Z. Tung, et al., "REALM: Retrieval-Augmented Language Model Pre-Training," *ICML*, 2020.

[21] D. Jin, et al., "What Disease Does This Patient Have? A Large-Scale Open-Domain Medical QA Dataset," *EMNLP*, 2021.

[22] N. Holzenberger, et al., "LegalBench: A Benchmark for Legal Reasoning in Large Language Models," *arXiv:2308.11462*, 2023.

[23] Z. Chen, et al., "FinQA: Numerical Reasoning Over Financial Data," *EMNLP*, 2021.

[24] Amazon Web Services, "Amazon Bedrock Documentation," AWS, 2023.

[25] Amazon Web Services, "Amazon OpenSearch Serverless Documentation," AWS, 2023.

[26] T. Zhang, V. Kishore, F. Wu, et al., "BERTScore: Evaluating Text Generation With BERT," *ICLR*, 2020.

[27] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *ACL Workshop*, 2004.

[28] C. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.