BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

ENTERPRISE DECISION INTELLIGENCE WITH RETRIEVAL-AUGMENTED GENERATION: A CONTEXT-AWARE FRAMEWORK FOR GROUNDED BUSINESS INSIGHTS

[1]Mansi Sachin Dusunge, [2]Dr. Vishal Puranik

[1]Department of Artificial Intelligence & Data Science, [2]Department of Artificial Intelligence & Data Science Engineering
[1,2]Ajeenkya DY Patil School of Engineering, Pune, India
[1]mansidusunge1@gmail.com, [2]vishalpura@gmail.com

Abstract

Organizations store decision-relevant information across financial reports, sales records, customer feedback, policy documents and market research, yet decision-makers often cannot retrieve the right facts quickly, and general-purpose AI chatbots answer from pre-trained knowledge that is neither current nor organization-specific. Retrieval-Augmented Generation (RAG) addresses this by retrieving relevant passages from a knowledge source before a large language model (LLM) generates its answer, which improves factual grounding and reduces hallucination. This paper reviews twelve studies spanning the retrieval-augmented generation lineage (RAG, REALM, RETRO, DPR, FiD, ColBERT), a recent survey of RAG techniques, a survey of LLM hallucination, and business-decision-support literature. It compares their methods and limitations, identifies gaps specific to enterprise deployment, comprising limited enterprise focus, multi-source data integration difficulty, residual hallucination, scalability and weak business-context awareness, and proposes a conceptual RAG-based Business Decision Assistant architecture: a query-processing layer, a dense retriever over an enterprise document index, an augmentation layer that fuses retrieved evidence into the prompt, and a generator that returns cited, confidence-flagged answers. The paper describes the architecture, the data flow, applicable governance and evaluation considerations, and a research agenda. The contribution is a literature-grounded design; it has not been implemented or benchmarked.

Keywords: Retrieval-Augmented Generation, large language models, business decision support, dense retrieval, enterprise knowledge management, hallucination mitigation, vector search

I. Introduction

A finance manager who wants last quarter's regional churn figures, a policy clause, and a competitor's pricing move in one answer cannot get all three from a spreadsheet, a PDF and a general chatbot separately without losing time. Large language models (LLMs) can write fluent, structured answers, but on their own they answer from parameters fixed at training time, so they cannot see an organization's current financial

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

reports, sales data, customer feedback or policies, and they are prone to hallucination, producing fluent but incorrect content [8].

Retrieval-Augmented Generation (RAG) addresses this by retrieving relevant passages from an external knowledge source and conditioning the LLM's generation on that retrieved evidence [1]. This idea has a substantial lineage, latent-retriever pretraining [3], retrieval over trillion-token corpora [2], dense passage retrieval [4], decoder-side passage fusion [5], and late-interaction retrieval [6], and has since been organized into a broader RAG paradigm with naive, advanced and modular variants [7]. Applying this lineage to enterprise decision support, rather than open-domain question answering, is comparatively recent and less thoroughly evaluated [9], [10].

This paper makes three contributions. First, it reviews twelve studies (2020–2025) spanning the RAG/retrieval lineage, RAG surveys, LLM hallucination, and business-decision-support literature, comparing their methods and limitations (Section III). Second, it derives research gaps specific to enterprise RAG deployment (Section IV). Third, it proposes a conceptual architecture for an AI-Powered Business Decision Assistant, comprising query processing, retrieval, augmentation and generation layers, together with governance and evaluation considerations (Sections V–VII). The review is purposive rather than exhaustive; the proposed architecture is a design contribution, not an implemented or benchmarked system.

II. Problem Statement and Objectives

A. Problem Statement

To study an AI-powered business decision assistant using Retrieval-Augmented Generation that integrates large language models with enterprise knowledge sources to deliver accurate, real-time, and context-aware business insights while minimizing AI hallucination and ensuring secure information access.

B. Objectives

1) Study the concept of Retrieval-Augmented Generation and large language models.
2) Understand the role of AI-powered decision assistants in business organizations.
3) Review the use of RAG for retrieving relevant organizational information.
4) Identify benefits, challenges and limitations of RAG-based business decision support.
5) Discuss the future scope and applications of AI-powered business decision support systems.

III. Related Work

Twelve studies published between 2020 and 2025 were selected from arXiv, ACL/EMNLP/SIGIR/ICML proceedings, Elsevier and IEEE venues to cover four strands: the retrieval-augmented generation lineage, surveys of RAG technique and LLM hallucination, and business-decision-support literature. Findings below are reported as stated by each source's authors; groupings and drawbacks are this review's own reading.

A. The Retrieval-Augmented Generation Lineage

Lewis et al. [1] introduce RAG itself: a pre-trained sequence-to-sequence generator is combined with a dense vector index of Wikipedia accessed through a learned neural retriever, and the model is fine-tuned

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

end-to-end. RAG sets a new state of the art on three open-domain QA benchmarks and produces more specific and factual language than a parametric-only baseline. The retrieval corpus in the original work is a fixed Wikipedia dump, and the method is evaluated on open-domain QA rather than enterprise data.

Guu et al. [3] propose REALM, which pre-trains a language model jointly with a latent knowledge retriever, backpropagating through a retrieval step over millions of documents using masked-language-modelling as the sole training signal. REALM outperforms prior explicit- and implicit-knowledge baselines by 4–16% absolute accuracy on open-domain QA and adds interpretability, since retrieved documents can be inspected. The retriever is trained unsupervised and is not evaluated on business or enterprise corpora.

Borgeaud et al. [2] present RETRO, which conditions an autoregressive LLM on document chunks retrieved, by local similarity, from a two-trillion-token database, using a frozen BERT retriever and a chunked cross-attention mechanism. RETRO matches GPT-3 and Jurassic-1 quality on the Pile while using 25x fewer parameters, and pre-trained transformers can be retrofitted with retrieval rather than retrained from scratch. The retrieval database is general web/book/code text, not organizational documents.

Karpukhin et al. [4] introduce Dense Passage Retrieval (DPR), replacing sparse TF-IDF/BM25 retrieval with dense embeddings learned by a dual-encoder trained on a modest number of question-passage pairs. DPR outperforms Lucene-BM25 by 9–19% absolute in top-20 retrieval accuracy and establishes new state-of-the-art results on several open-domain QA benchmarks. DPR retrieves passages only; it does not generate an answer or provide end-to-end decision support on its own.

Izacard and Grave [5] propose Fusion-in-Decoder (FiD), which encodes each retrieved passage independently and fuses them only in the decoder's cross-attention, letting a generative model scale to many more supporting passages than single-passage RAG formulations while remaining computationally tractable. FiD improves answer accuracy on open-domain QA benchmarks over prior retrieval-generation combinations. The evaluation, like RAG's, is confined to question answering rather than enterprise applications.

Khattab and Zaharia [6] present ColBERT, a late-interaction retrieval architecture that encodes queries and documents independently with BERT and then applies a lightweight token-level interaction step, allowing document representations to be pre-computed offline. ColBERT retains much of the effectiveness of full cross-attention rerankers at a fraction of the query-time cost and can retrieve directly from a full collection using vector-similarity indexes. The method is evaluated on MS MARCO and TREC-CAR passage search, not on organizational knowledge bases.

B. Surveys of RAG Technique and LLM Hallucination

Gao et al. [7] systematically survey the RAG literature and organize it into Naive, Advanced and Modular RAG paradigms, covering retrieval, generation and augmentation techniques along with evaluation frameworks and open resources. They identify multi-modality expansion and RAG-infrastructure maturity as open research directions. The survey is technique-centred and does not focus on any single application domain such as enterprise decision support.

Huang et al. [8] survey hallucination in LLMs, proposing a taxonomy of hallucination types, analysing contributing factors across the model lifecycle, and reviewing detection methods and benchmarks. They note that retrieval-augmented approaches are among the most effective mitigations because they ground

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

generation in external, verifiable evidence, which is the mechanism this paper's proposed assistant relies on. The survey covers general-purpose LLM use and does not analyse business-specific hallucination risk.

C. Business Decision Support with LLMs

Handler et al. [9], writing in the International Journal of Information Management, argue that LLMs are a general-purpose decision-support technology rather than the narrow, bespoke tools traditional decision-support research has studied, and use Herbert Simon's classic three-stage model (intelligence, design, choice) to organize new research questions LLMs raise at each stage of organizational decision-making. The paper is conceptual and proposes a research agenda rather than a system.

Seabra et al. [10] describe "Pergunte ao BI" ("Ask the BI"), a conversational interface that uses LLMs to support business-intelligence decision-making, demonstrated at the Brazilian Symposium on Databases (SBBD) 2025. The system lets non-technical users query BI dashboards and data warehouses in natural language rather than writing SQL or navigating dashboard menus. As a demonstration paper, it reports system design rather than a controlled evaluation.

Raiaan et al. [11] review LLM architectures, applications, taxonomies and open challenges in IEEE Access, covering enterprise and business use cases alongside healthcare, education and other domains, and catalogue open issues including hallucination, bias, computational cost and data privacy. The review is broad rather than business-decision-specific.

Cheung [12] runs real-world experiments with four LLMs (GPT-3, Google Bard, NeevaAI, Claude Instant) on core business tasks such as project-planning reference suggestions, reporting a quantified analysis of their usefulness and limitations and highlighting the need for prompt engineering in practical deployment. The evaluation is limited to five reference business tasks and does not test a retrieval-augmented configuration.

Table I: Comparative Summary of Reviewed Literature

Ref.	Strand	Approach	Key finding	Limitation
[1]	RAG lineage	Seq2seq generator + dense Wikipedia index	SOTA on 3 open-domain QA benchmarks	Fixed Wikipedia corpus; not enterprise
[3]	RAG lineage	REALM: latent retriever pre-trained with LM	+4–16% accuracy; adds interpretability	Unsupervised retriever; no business data
[2]	RAG lineage	RETRO: retrieval over 2T-token database	Matches GPT-3 with 25x fewer params	General text corpus, not org documents
[4]	RAG lineage	DPR: dual-encoder dense retrieval	+9–19% over BM25 retrieval	Retrieval only; no generation

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Ref.	Strand	Approach	Key finding	Limitation
[5]	RAG lineage	FiD: decoder-side passage fusion	Higher QA accuracy, scales to more passages	QA-focused, not enterprise
[6]	RAG lineage	ColBERT: late-interaction retrieval	Near cross-attention accuracy, low cost	Evaluated on MS MARCO/TREC-CAR
[7]	RAG survey	Naive/Advanced/Modular RAG taxonomy	Systematises retrieval, generation, augmentation	Technique survey, no domain focus
[8]	Hallucination survey	Taxonomy of causes, detection, mitigation	Retrieval grounding is a key mitigation	General-purpose, not business-specific
[9]	Business decision support	Simon's 3-stage model applied to LLMs	LLMs are general-purpose decision-support tech	Conceptual; no system
[10]	Business decision support	Conversational BI interface (LLM)	Natural-language access to BI dashboards	Demo paper; no controlled study
[11]	LLM review	Architectures, taxonomies, open issues (IEEE Access)	Catalogues hallucination, bias, cost, privacy issues	Broad, not decision-support-specific
[12]	Business LLM evaluation	4 LLMs tested on business tasks	Quantified usefulness and limitations	5 reference tasks; no RAG tested

As Fig. 1 shows, the retrieval/RAG lineage accounts for exactly half of the reviewed base, which is consistent with Section III's framing that core retrieval-generation technique is more mature than its business-decision application. Fig. 2 shows the review is weighted toward 2020 (the year RAG, REALM and DPR were all published) and 2024–2025 (the year the business-decision-support literature emerged), with comparatively fewer intervening sources – itself a reflection of the gap this paper identifies in Section IV.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

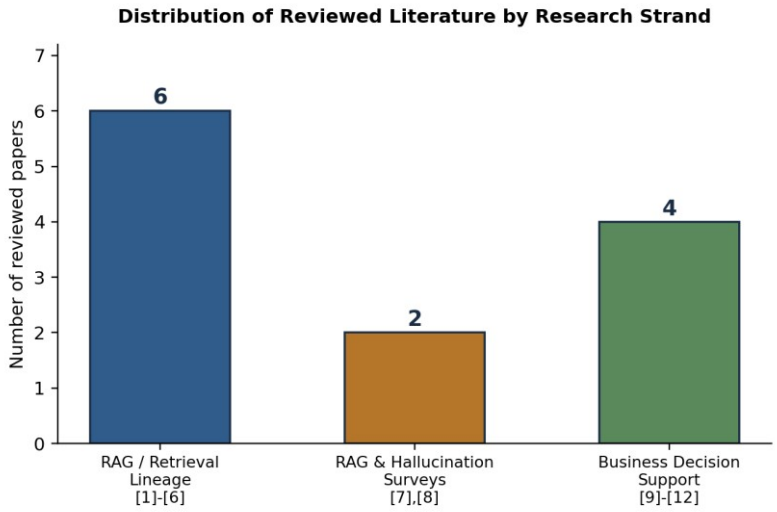


Fig. 1: Distribution of the twelve reviewed papers across the three literature strands tallied from Table I.

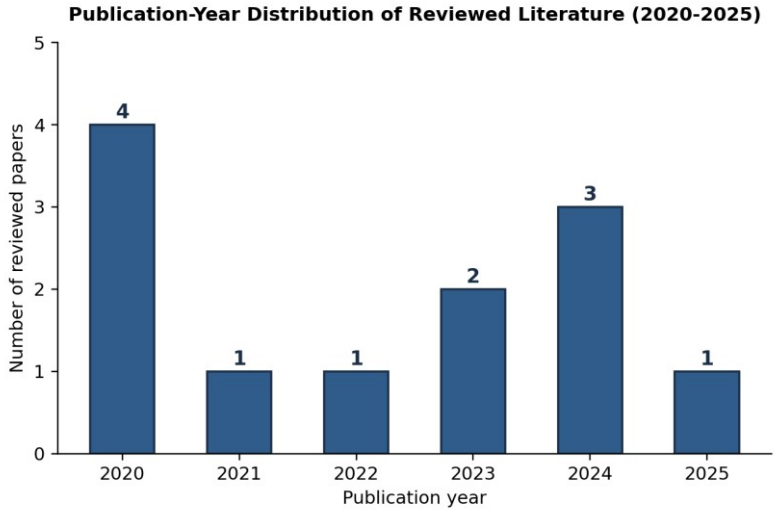


Fig. 2: Publication-year spread of the reviewed literature, confirming the 2020–2025 recency window claimed in Section III.

IV. Research Gap

Reading the twelve sources together surfaces five gaps that motivate the architecture in Section V.

G1. Limited enterprise focus. Most RAG systems in the core lineage are built and evaluated for general open-domain question answering [1], [3], [4], [5], [6], not business-specific decision-making.

G2. Data-integration difficulty. Enterprise information is spread across PDFs, databases, reports and spreadsheets in heterogeneous formats, and the reviewed retrieval methods assume a single, pre-processed text corpus [1], [2], [4].

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

G3. Residual hallucination. RAG reduces but does not eliminate hallucination; when relevant information is missing or incomplete, generation can still be inaccurate or misleading [8].

G4. Scalability. Retrieval methods validated on benchmark corpora [4], [5], [6] are not necessarily optimized for large-scale, constantly updated organizational knowledge bases.

G5. Weak business-context awareness. General-purpose LLM decision-support studies [9], [11], [12] and BI-specific demonstrations [10] each address part of the problem, but no reviewed source combines retrieval grounding with an architecture purpose-built for organization-specific, secure, context-aware business decisions.

Main gap: current solutions are not fully capable of providing fast, accurate, real-time, organization-grounded insight for business decision-making in dynamic environments; the retrieval, generation and hallucination-mitigation pieces exist separately in the literature but have not been combined into one architecture purpose-built for enterprise decision support.

V. Proposed System Architecture

A. Design Principles

The architecture responds directly to the gaps in Section IV: it targets business queries rather than open-domain QA (G1), ingests heterogeneous enterprise formats (G2), grounds every answer in cited evidence with a confidence flag (G3), separates retrieval from generation so the index can scale independently (G4), and keeps a business-context layer between the raw retriever and the LLM (G5).

- **Grounding before generation:** no answer is produced without first retrieving supporting passages, following the RAG pattern in [1] and the fusion approach in [5].
- **Dense, semantic retrieval:** queries and documents are embedded and matched by similarity, as in DPR [4] and ColBERT [6], rather than relying on keyword search alone.
- **Access-controlled retrieval:** the retriever only searches documents the querying user is authorised to see, addressing the secure-access requirement in the problem statement.
- **Cited, confidence-flagged output:** every generated answer carries its source passages and a confidence indicator, directly targeting the hallucination risk documented in [8].

B. Architecture Overview

Fig. 3 shows four layers. The Query Processing layer parses user intent and rewrites the query for retrieval, and applies an access-control check before anything is retrieved. The Retriever, following the dense-embedding approach of DPR [4] and the late-interaction efficiency of ColBERT [6], searches a vector index built over enterprise knowledge sources: financial reports, sales data, CRM records, policy documents and market research. The Augmentation layer fuses the top-k retrieved passages with the original query into the LLM's context, following the decoder-side fusion strategy of FiD [5] rather than single-passage conditioning. The Generator, an LLM, produces a grounded answer, and the final response is returned with source citations and a hallucination/confidence check, addressing G3.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

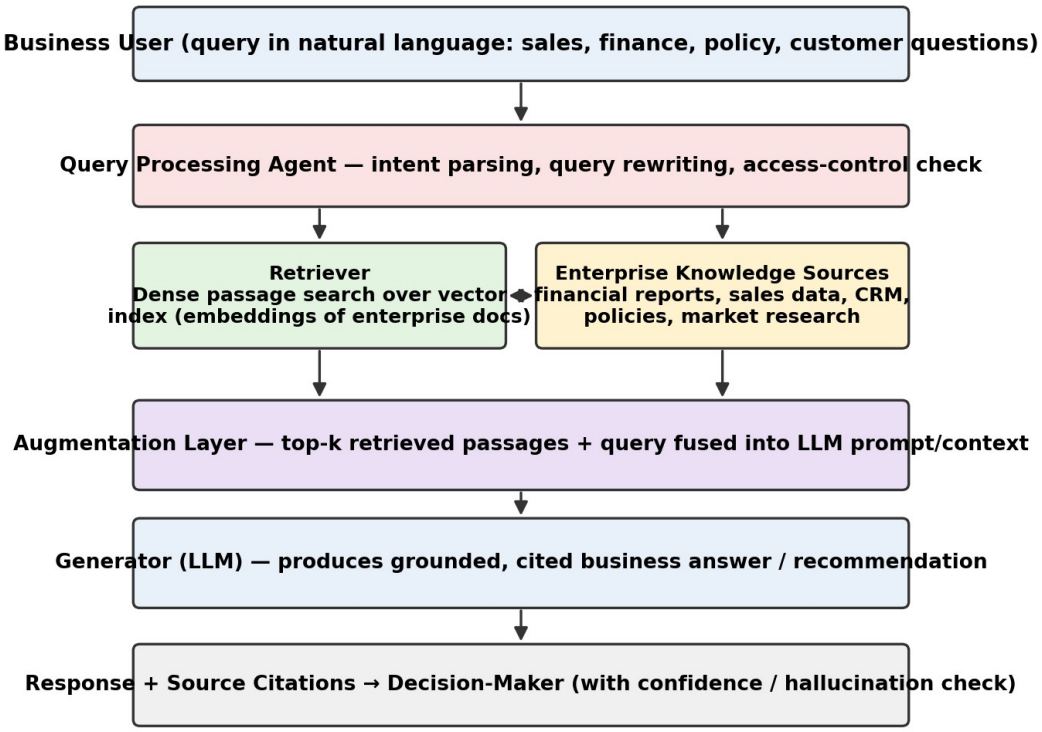


Fig. 3: Architecture of the AI-Powered Business Decision Assistant.

Table II: Architecture Components

Component	Role	Basis in reviewed work
Query Processing	Intent parsing, access control	New layer; not present in [1]– [6]
Retriever	Dense passage search	DPR [4], ColBERT [6]
Augmentation	Fuse passages into LLM context	FiD [5], RAG [1]
Generator	Produce grounded, cited answer	RAG [1], hallucination mitigation [8]

C. Data Flow

A single query proceeds as follows (Fig. 4): (1) the user submits a business question in natural language; (2) the query is embedded into the same vector space as the indexed documents; (3) the retriever fetches the top-k relevant chunks; (4) those chunks are fused with the query into the LLM's context; (5) the LLM generates an answer grounded in the retrieved evidence; (6) the answer is returned with citations, and low-confidence answers, for example where retrieval similarity is weak, are flagged for human review rather

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

than presented as fact, which is the mechanism the framework uses to address the hallucination risk in [8] specifically for business use.

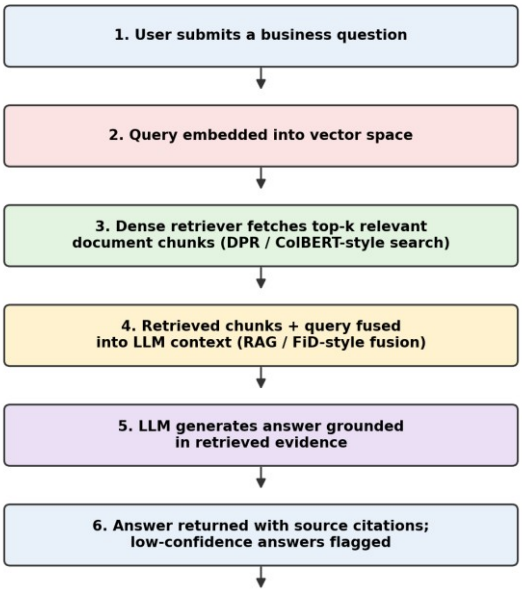


Fig. 4: End-to-end RAG pipeline for a business query.

VI. Comparative Analysis

Table III compares the proposed assistant with the reviewed retrieval-generation methods and business-decision-support studies on the dimensions that define the gaps in Section IV. Entries for existing work reflect what the cited papers report; the proposed-assistant column is design intent, not a measured result.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Table III: Comparison Across Dimensions

Dimension	RAG lineage [1]– [6]	Business decision studies [9], [10], [12]	Proposed assistant
Enterprise-document retrieval	No (open-domain corpora)	Partial (BI dashboards only [10])	Yes (intended)
Multi-format data integration	No	Not addressed	Yes (intended)
Hallucination mitigation	Partial (retrieval grounding)	Not addressed	Yes, with confidence flag
Access-controlled retrieval	No	Not addressed	Yes (intended)
Empirical business validation	No	Limited [10], [12]	None yet

The last row is the important caveat: like the frameworks it builds on, the proposed assistant's advantages are architectural claims until it is implemented and tested on real organizational data. Fig. 5 makes the same point visually – the bottom-right cell is red, not green, by design: no empirical validation exists yet for the proposed assistant, and the figure does not overstate this.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

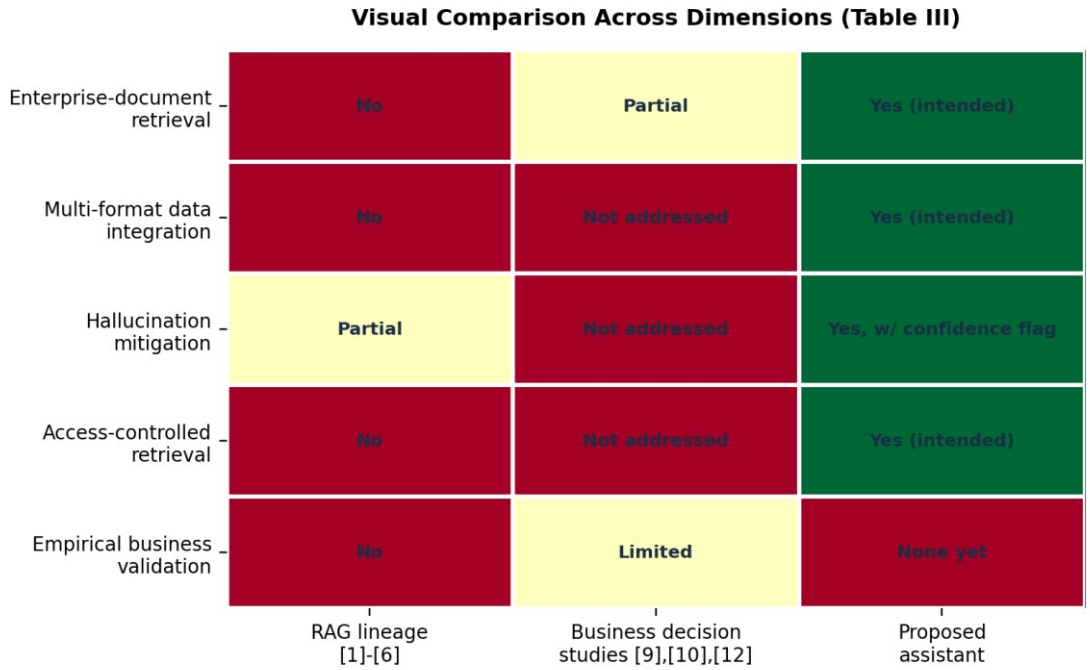


Fig. 5: Visual rendering of Table III; green marks full/intended capability, yellow partial or limited coverage, red the absence of the capability in that body of work.

VII. Ethical and Practical Considerations

Business documents indexed for retrieval, financial reports, customer feedback, HR policies, can contain commercially sensitive or personal data, so the access-control check in the Query Processing layer is a design requirement, not an optional feature; retrieval must respect the same permissions the underlying documents already carry. Hallucination remains a risk even with retrieval grounding [8], so answers used for consequential decisions should be treated as decision support rather than an automatic final answer, consistent with the framing in [9] that LLMs assist stages of decision-making rather than replace the decision-maker's judgement. Any organization deploying such a system should audit retrieval logs, since queries and retrieved passages themselves can reveal what an employee was investigating.

VIII. Evaluation Plan

Because the architecture has not been implemented, Table IV sets out the questions and measures a future study would need.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Table IV: Proposed Evaluation Plan

Question	Measure	Basis
Is retrieval accurate?	Top-k retrieval accuracy	DPR/ColBERT method [4], [6]
Are answers grounded?	Citation coverage; factual-consistency check	Hallucination benchmarks [8]
Is it useful for business tasks?	Task-based evaluation on real business queries	Approach used in [12]
Is access controlled?	Audit of retrieval logs against permissions	Design requirement (Sec. VII)

IX. Conclusion and Future Work

This review examined twelve studies spanning the retrieval-augmented generation lineage [1]– [6], surveys of RAG technique and LLM hallucination [7], [8], and business-decision-support literature [9]– [12], and found that each strand addresses part of the enterprise decision-support problem without addressing all of it: strong retrieval and generation methods exist for open-domain QA, and decision-support and BI literature identifies the business need, but no reviewed source combines both into one architecture aimed at organization-specific, secure, real-time business decisions. Building on this gap, we proposed a conceptual RAG-based Business Decision Assistant with query-processing, retrieval, augmentation and generation layers. The contribution is a literature-grounded design and research agenda; it has not been implemented or benchmarked.

Future work should prioritise a prototype connected to real (or realistically simulated) enterprise documents, evaluation of retrieval accuracy and hallucination rate on business-specific queries following the metrics in Table IV, and a usability study with actual business decision-makers.

Acknowledgment

The author thanks Dr. Vishal Puranik for guidance during the literature survey and design stages of this technical seminar.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Proc. NeurIPS, 2020, arXiv:2005.11401.

[2] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, et al., “Improving Language Models by Retrieving from Trillions of Tokens,” in Proc. ICML, PMLR vol. 162, pp. 2206–2240, 2022, arXiv:2112.04426.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

[3] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in Proc. ICML, PMLR vol. 119, pp. 3929–3938, 2020, arXiv:2002.08909.

[4] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, pp. 6769–6781, 2020, arXiv:2004.04906.

[5] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in Proc. EACL, 2021, arXiv:2007.01282.

[6] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in Proc. ACM SIGIR, pp. 39–48, 2020, doi: 10.1145/3397271.3401075.

[7] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2023.

[8] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," arXiv:2311.05232, 2023.

[9] A. Handler, K. R. Larsen, and R. Hackathorn, "Large Language Models Present New Questions for Decision Support," Int. J. Inf. Manage., vol. 79, 2024, doi: 10.1016/j.ijinfomgt.2024.102865.

[10] A. Seabra, C. Cavalcante, G. Medeiros, J. Nepomuceno, V. Sallenave, N. Ruberg, and S. Lifschitz, "Pergunte ao BI: Uma Interface Conversacional para Apoio a Tomada de Decisao usando Modelos de Linguagem," in Anais SBBD 2025 (Demonstracoes e Aplicacoes), Porto Alegre, Brazil, 2025, pp. 142–147, doi: 10.5753/sbbd_estendido.2025.247794.

[11] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," IEEE Access, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[12] M. Cheung, "A Reality Check of the Benefits of LLM in Business," arXiv:2406.10249, 2024.