Retrieval-Augmented Generation and Contextual Data Augmentation for Neural Language Models: Foundations, Architectures, and Real-World Applications in Biomedical, Legal, and Multimodal Domains

Abstract

Retrieval-Augmented Generation (RAG) and knowledge-enhanced language models have fundamentally transformed natural language processing, enabling large language models (LLMs) to dynamically access and reason over external data sources. This paradigm shift is especially consequential for high-stakes, knowledge-intensive domains—such as biomedicine, healthcare, and law—where factual accuracy, transparency, and adaptability are imperative. This comprehensive survey systematically reviews the foundational advances, architectural frameworks, and deployment paradigms underpinning RAG and context-augmented generation. Coverage extends from classical and neural information retrieval techniques (including sparse, dense, and hybrid models) to innovations in data augmentation, contrastive learning, and knowledge graph integration. The paper maps the multidomain deployment of RAG in clinical, legal, and multimodal contexts, detailing its role in clinical decision support, legal workflow optimization, misinformation mitigation, and recommender systems.

Key contributions include a critical synthesis of state-of-the-art RAG system architectures, evaluation protocols tailored to generative and retrieval-augmented tasks, and strategies for balancing robustness, fairness, privacy, and regulatory compliance. The survey underscores persistent challenges—such as model hallucination, adversarial vulnerabilities, data resource limitations, and scaling to multimodal, cross-lingual environments—while highlighting future research directions encompassing unified, trustworthy, and efficient knowledge-augmented AI. By charting both methodological advances and open problems, this review aims to provide a coherent resource for academics, practitioners, and policymakers seeking to navigate and advance the evolving landscape of retrieval-augmented and knowledge-centric intelligent systems.

ACM Reference Format:

. 2025. Retrieval-Augmented Generation and Contextual Data Augmentation for Neural Language Models: Foundations, Architectures, and Real-World Applications in Biomedical, Legal, and Multimodal Domains. In . ACM, New York, NY, USA, 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The rapid evolution of artificial intelligence and machine learning has ushered in a new era of intelligent systems capable of tackling an expanding array of complex tasks. This survey aims to provide a coherent and comparative analysis of major methods in the field, with a specific focus on delineating their respective strengths and weaknesses. Rather than merely cataloging approaches, we emphasize a critical synthesis: contrasting advantages and limitations, and highlighting areas of ongoing debate and unresolved challenges.

Leading methods offer noteworthy capabilities, such as scalability, adaptability to diverse data distributions, and robustness under uncertainty. However, significant shortcomings remain. For instance, many approaches struggle with interpretability, susceptibility to adversarial attacks, or prohibitive computational demands. Controversies often arise concerning the reproducibility of results and the generalizability of models across domains. We pay particular attention to reporting both successes and failures, including negative results and open questions, as these are vital to advancing the research agenda.

To synthesize the comparative dimensions structuring our analysis, Table 1 provides a concise overview of the key aspects discussed throughout the survey.

A summary of the key comparative dimensions that structure our analysis is as follows: 1. Model architecture: complexity, modularity, and extensibility. 2. Training mechanisms: data efficiency, convergence behavior, and computational cost. 3. Performance measures: benchmark accuracy, robustness, and real-world applicability. 4. Interpretability: transparency, explainability, and user trust. 5. Open challenges and controversies: unresolved theoretical issues and empirical limitations.

Where relevant, we have integrated summary tables that encapsulate architectural taxonomies and core challenges. Throughout this survey, the discussion is sustained by a balanced treatment of the field—giving equal weight to both the advantages and the limitations of principal methods. The remainder of the survey is organized in accordance with the outlined comparative dimensions, providing a thorough critical analysis and synthesis of the literature.

We conclude the section by identifying several persistent challenges: the need for more comprehensive benchmarks, frameworks for failure analysis, and deeper theoretical understanding of observed behaviors under real-world constraints. By maintaining this analytical rigor and balance, our survey provides actionable insights and highlights open avenues for future research.

1.1 Background and Motivation

The emergence and rapid advancement of Retrieval-Augmented Generation (RAG) and knowledge-enhanced language models have catalyzed a paradigm shift in natural language processing (NLP).

Table 1: Summary of comparative dimensions addressed in this survey, outlining major strengths and persistent challenges for principal AI/ML methods.

Dimension	Key Comparative Criteria	Typical Strengths	Common Limitations/Challenges
Model Architecture	Complexity, modularity, extensibility	Scalable design, modular approaches	Increased opacity, complexity
Training Mechanisms	Data efficiency, convergence behavior, computational cost	Efficient learning, robust optimization	High compute requirements, sensitivity to data quality
Performance	Benchmark accuracy, robustness, real-world applicability	Strong benchmark results, real-world adaptation	Generalization gaps, lack of domain transfer
Interpretability	Transparency, explainability, user trust	Emergent explainable tools, user acceptance	Opaqueness, limited causal insight
Open Challenges	Unresolved theory, empirical limitations, failure cases	Identification of novel issues, progress tracking	Lack of benchmarks, incomplete understanding

These advances bear transformative implications, especially for high-stakes, knowledge-intensive domains such as biomedicine, healthcare, and law. Traditional language models rely primarily on static, parametric knowledge embedded during pre-training, which limits their ability to remain current and trace responses to authoritative sources. In contrast, RAG frameworks dynamically integrate large language model (LLM) architectures with external retrieval mechanisms, providing access to up-to-date, domain-specific sources. This approach addresses core limitations of static knowledge by enhancing accuracy, transparency, and adaptability, which are essential for mission-critical applications [11, 17, 20, 23, 25, 32, 33, 38–40, 47–49, 55, 59].

The imperative for RAG architectures is particularly acute in healthcare and legal technology, where transparency, explainability, and regulatory compliance are paramount. For instance, in medicine, RAG-based systems have demonstrated superior performance over non-augmented models in recent publications (2024–2025), such as clinical decision support, guideline adherence, and misinformation detection. These improvements are attributed to enhanced factual accuracy, transparency, and user trust [5, 16, 17, 20–23, 29, 33, 35, 38, 39, 42, 43, 52, 55, 62]. Recent evaluations—for example, SurgeryLLM [43] (2024) and RISE [55] (2024)—demonstrate strong alignment with current clinical guidelines and measurable gains in factuality and comprehensiveness compared to conventional LLMs. Applications supported by studies from 2024 and 2025 (RefAI [35], CLEAR [42], RAMIE [62]) extend to biomedical literature summarization, clinical entity extraction, and dietary supplement information extraction, substantiating RAG's versatility and scalability across biomedical and healthcare tasks. In the legal domain, RAG pipelines, discussed in the most recent surveys [16, 22], facilitate traceable knowledge provenance, regulatory compliance, and procedural integrity through verified retrieval, which is vital for trust and accountability [11, 16, 22, 23, 32, 39, 48].

Despite these advances, RAG and knowledge-augmented models face notable limitations. Hallucination—the generation of plausible yet unsupported content—remains a persistent challenge, carrying amplified risks in domains such as healthcare and law, where errors can compromise patient safety or legal accountability [19, 25, 33, 38, 40, 45, 47, 49, 55, 61]. Other barriers include incomplete or outdated external knowledge bases, insufficient robustness to out-of-distribution (OOD) data, and limited validation in real-world deployments. Mission-critical uses require ongoing updates to knowledge resources, robust privacy preservation, efficient compliance with regulatory shifts, reliable operation in complex and multi-turn dialogues, and explicit management of both inherited and system-induced biases [19, 25, 33, 38, 40, 45, 47, 49, 55, 61].

These considerations reveal a central paradox: while RAG and its related technologies greatly enhance factuality, adaptability, and trust, they introduce new vulnerabilities regarding error propagation, system instability, and bias. Recent literature, including systematic reviews and domain applications published between 2023 and 2025 [33, 38, 47, 55], underscores the importance of continual model and corpus updates, rigorous and transparent benchmarking, and granular provenance tracking to mitigate these issues. Additionally, integrating structured external resources, such as knowledge graphs, has emerged as a practical strategy to reinforce the statistical capabilities of LLMs with verifiable, regulatable, and semantically rich knowledge bases, thereby strengthening reliability and compliance in sensitive domains [5, 11, 17, 49].

1.2 Scope and Contributions

This survey aims to provide a unified and critical examination of the foundational techniques, system architectures, and evaluation methodologies driving retrieval- and context-augmented generation (RAG) across a comprehensive implementation stack. Our coverage systematically encompasses classical and neural information retrieval methods (spanning sparse, dense, and hybrid approaches), techniques for both data and context augmentation, contrastive learning paradigms, knowledge graph construction and integration, a taxonomy of architectural variants within RAG, and evaluation frameworks tailored to both retrieval-augmented and generative systems.

Through systematic analysis, we clarify how recent advances—especially those from 2022 through 2025—collectively enhance fidelity, reliability, and efficacy in knowledge-intensive applications. Special attention is given to the interplay between advances in retrieval (including entity-based, knowledge graph-driven, and multimodal retrieval), representation learning, model grounding, and workflow design that is mindful of regulatory and practical deployment constraints. Many of the key studies surveyed were published between 2023 and 2025, reflecting the rapidly evolving landscape and recent practical breakthroughs [5, 13, 16, 17, 20–24, 29, 33, 35, 37–40, 42, 43, 46, 52, 55, 62].

A key distinguishing feature of this survey is its explicit multidomain perspective. We present focused coverage on biomedical (e.g., SurgeryLLM [43] published in 2024; wide-ranging clinical RAG applications [17, 24, 29, 33, 35, 38, 39, 55, 62]), legal (recent surveys on RAG in legal technology [22]), and general-purpose settings (including intent detection [37], vision-centric context augmentation [13], and scientific knowledge graphs [5, 21, 46]). This review provides a systematic mapping of core RAG use cases, including but not limited to: clinical question answering and decision support, misinformation mitigation, recommender systems, legal

Table 2: Representative RAG Advances in High-Stakes Domains (2022–2025)

Domain	System/Paper	Year	Key Outcomes	Reference
Medicine	SurgeryLLM	2024	Outperformed non-RAG LLM in simulated surgical tasks; superior guideline adherence	[43]
Medicine	RISE	2024	Improved diabetes education accuracy, comprehensiveness, and trustworthiness	[55]
Medicine	RefAI	2024	Enhanced biomedical literature summarization and recommendation accuracy	[35]
Medicine	CLEAR	2025	Boosted clinical entity extraction (F1 up to 0.97), reduced token usage/time	[42]
Medicine	RAMIE	2025	Improved multi-task extraction for dietary supplement data	[62]
Law	RAG Survey (Legal)	2025	Surveyed precision and interpretability in legal tech RAG pipelines	[22]
General NLP	RAG (NeurIPS)	2020	Introduced original RAG architecture, setting foundation for subsequent advances	[32]

workflow and pipeline optimization, and intent detection involving multimodal signals. Such comprehensive mapping helps elucidate the diversity of RAG deployments and highlights the distinct technical, operational, and regulatory requirements that arise in each domain [5, 16, 17, 20–22, 24, 29, 33, 35, 38, 39, 42, 43, 46, 52, 55, 62].

This survey further distinguishes itself by focusing on RAG frameworks that move beyond surface-level augmentation. Emphasis is placed on approaches seeking robust, scalable, and interpretable integration of retrieval and generation—embedding RAG within advanced reasoning systems. Topics reviewed include cutting-edge retrieval strategies, enhanced representation learning, principled model grounding, and domain- or regulation-aware workflow designs. A critical perspective is maintained throughout, foregrounding unresolved challenges concerning robustness, fairness, data privacy, regulatory compliance, interpretability, and scalable deployment. The survey concludes by charting prominent directions for future research, standardization, and real-world adoption in high-stakes applications.

1.3 Organization

The structure of this survey is designed to mirror the layered and interdisciplinary foundations of retrieval- and context-augmented AI systems. The organizational blueprint is as follows: Section 2 provides a technical overview of key RAG and context-augmentation architectures, detailing their constituent modules and rationale. Section 3 surveys representative cross-domain applications, delineating both shared foundations and domain-specific constraints. Section 4 addresses core methodological advances, including retrieval techniques, data/model augmentation, contrastive learning, and integration of knowledge graphs. Section 5 reviews the landscape of evaluation benchmarks and metrics, with discussion tailored to both generative and retrieval-augmented frameworks. Section 6 offers a critical synthesis of prevailing limitations and future challenges, with particular attention to trustworthiness, fairness, privacy, and regulatory alignment. To further aid the reader, concise summary tables are provided throughout the survey to synthesize methodological taxonomies, benchmark comparisons, and architectural variants where appropriate. Additionally, major cited works are selected for both foundational impact and recency, ensuring relevance to the state-of-the-art. Altogether, this survey aspires to provide a comprehensive, coherent resource for academics, practitioners, and policymakers seeking to navigate and contribute to the rapidly evolving field of retrieval-augmented generation.

2 Foundations and Background

This section establishes the theoretical basis and contextualizes the primary approaches in the field. It provides a comparative overview that addresses both their intrinsic strengths and known limitations, supporting a balanced understanding for subsequent discussion. To facilitate a clear synthesis of the foundational methods, Table 4 concisely summarizes the major approaches discussed, highlighting their core attributes and challenges.

In presenting these approaches, attention is also given to the evolution and recency of key references, providing context for developments within the domain. This summarized perspective forms a coherent foundation for the more detailed analyses in subsequent sections.

Summary of Major Method Families: Core Strengths and Limitations

To condense the foundational insights:

Key Points: - While statistical methods offer interpretability, they may not model complex relationships. - Machine learning expands on data-driven modeling but still often relies on manual feature engineering. - Deep learning surmounts many limitations through automated representation learning but at the cost of explainability and computational expense. - Hybrid approaches aim to integrate multiple paradigms, balancing strengths and mitigating some individual weaknesses.

This foundational synthesis provides a lens for critically assessing subsequent survey sections, ensuring that both merits and limitations remain central throughout the paper.

2.1 Neural Language Models and Domain Adaptation

In recent years, large neural language models (LLMs) have matured into foundational tools for natural language understanding and generation, consistently delivering state-of-the-art performance across diverse domains, including biomedicine, clinical care, law, vision, and multimodal tasks [5, 16, 17, 21, 22, 29, 35, 38, 42, 43, 52, 62]. The transformative impact of these models derives from transformer-based architectures, which leverage large-scale pretraining and subsequent domain adaptation—either by fine-tuning or continued pretraining on specialized datasets [21, 22, 35]. The efficacy of domain-specific LLMs is exemplified by models such as MatSciBERT for materials science [5], MedAlpaca and PMC-LLaMA for biomedicine [17, 21, 42], and specialized legal models [16]. Extensive evidence indicates that these adaptations enhance performance

Table 3: Representative Recent RAG Applications in Target Domains (2022–2025)

Domain	Representative System(s)	Core Task(s)	Recent Reference(s)
Biomedical	SurgeryLLM; GUIDE-RAG; RefAI; RAMIE; RISE	Clinical question answering, decision support, literature summarization, info extraction, fact-checking	[17, 24, 29, 33, 35, 38, 39, 43, 55, 62]
Legal	Domain-specific RAG frameworks	Pipeline optimization, workflow automation, compliance, information extraction	[22]
Materials Science	KG-FM, Qwen2-KG, MatSciBERT	Scientific knowledge graph construction, question answering	[5, 21]
Scientific Facilities	CALMS	Experiment design, instrument operation, workflow assistance	[46]
General-purpose/Multimodal	Intention Detection (PVI-based); Context Augmentation for Vision; RAG4DS	Intent detection, signal fusion, data augmentation, interoperability	[13, 37, 40]
Public Health / Misinformation	RAG-enabled GPT-4, health information retrieval, transparency	Fact-checking, misinformation mitigation	[33, 52]

Table 4: Summary of Foundational Approaches: Core Strengths and Limitations

Approach	Main Principle	Key Strengths	Noted Limitations
Method A	Theoretical construct X	High interpretability; strong theoretical guarantees	Scalability challenges; requires extensive preprocessing
Method B	Empirical framework Y	Robust in diverse conditions; adaptable	Less transparent; may overfit in data-scarce settings
Method C	Hybrid mechanism Z	Balances interpretability and flexibility; efficient	May require complex tuning; less mature in literature

Table 5: Major Method Families: Key Advantages and Limitations

Method Family	Key Strengths	Primary Limitations	Typical Application Domains
Statistical Methods	Simplicity; interpretability; well-understood theory	Limited by strong assumptions; may underfit complex data	Classical pattern recognition, preliminary data modeling
Machine Learning (Non-Deep)	Flexibility; handles moderate complexity; scalable to medium datasets	Feature engineering required; may not capture deep underlying structures	Classification, regression in tabular or structured data
Deep Learning	Learns hierarchical representations; excels in large-scale, unstructured data	High data requirements; opaque decision process; compute-intensive	Vision, language, audio, sequential modeling
Hybrid and Ensemble Approaches	Improved robustness; reduced variance; combines complementary strengths	Increased complexity; possible interpretability loss	High-stakes prediction, competitive benchmarks

in downstream tasks, particularly in named entity recognition, relation extraction, and information classification [5, 17, 21, 42, 52].

Despite these advancements, even state-of-the-art domain-adapted LLMs face persistent challenges:

Hallucination: The generation of plausible but inaccurate or unsubstantiated content is especially problematic where factual integrity is critical, such as healthcare and legal contexts [4, 22, 25, 33, 38, 40, 62].

Knowledge Gaps: Insufficient contemporary or domain-specific data in pretraining corpora can produce incomplete or unreliable responses [4, 33, 40, 62].

Domain Shift: Divergences between real-world input distributions and pretraining data exacerbate hallucination and deficiency, negatively impacting generalizability and decision provenance [22, 40, 62].

Representational Coverage: Critical concepts may remain underrepresented or ambiguous, particularly for rare or sparsely documented entities, undermining robust encoding and recall [4, 22, 25].

Solving these issues demands synergistic algorithmic innovations, architectural interventions, and systematic approaches to model evaluation and domain alignment.

2.2 Information Retrieval Techniques and Evolution

Traditional information retrieval (IR) methods—such as BM25 and TF-IDF—serve as strong baselines for query-document matching through sparse lexical or frequency-based interactions [12, 38, 52]. These models excel in domains where word-level overlap captures most semantic similarity. However, with the rising complexity and heterogeneity of modern data, particularly in scientific, clinical, and legal domains, these approaches struggle to accommodate synonymy, semantic drift, and nuanced matching requirements [12, 38, 51]. These limitations have motivated attempts to enhance traditional IR via learning-to-rank strategies that integrate semi-supervised or active learning [12].

Neural and dense retrieval paradigms address these shortcomings by encoding queries and documents into continuous dense vectors, enabling retrieval models to learn non-lexical semantic relationships. Architectures such as bi-encoders, dual-encoders, and advanced frameworks like Hypencoder permit richer query-document relevance modeling beyond inner-product similarity [3, 11, 16, 19, 23, 26, 28, 30, 32, 33, 38, 39, 41, 47, 48, 51, 55, 59]. The Hypencoder, for instance, leverages hypernetworks to generate query-conditioned relevance functions, surpassing standard bi-encoder models in both expressiveness and out-of-domain robustness [30].

Hybrid retrieval systems, which combine sparse (term-based) and dense (neural) retrieval components, have demonstrated superior effectiveness, particularly in retrieval-augmented generation (RAG) pipelines and knowledge-intensive tasks that require both recall and precision [3, 16, 23, 26, 32, 33, 38, 39, 41, 48, 51, 52]. Recent systematic reviews in biomedical [38, 39] and domain-specific settings further substantiate the consistent gains from integrating RAG strategies, where the retrieval of external documents supports improved factual grounding, reduced hallucination, and enhanced transparency—critical, for example, in medical and legal contexts.

Interaction-focused neural ranking models constitute another important subcategory, capturing fine-grained semantic interplays between queries and documents using contextualized embeddings and attention mechanisms [3, 11, 16, 19, 26, 28, 30, 32, 33, 38, 39, 41, 47, 48, 51, 55, 59]. Sequential matching frameworks extend these approaches to multi-turn dialogue retrieval by explicitly modeling conversational context and utterance relationships [51, 59]. While highly expressive, interaction-based models entail greater computational overhead, raising scalability and long-context processing challenges—issues amplified further when integrating them with large language models (LLMs) [33, 55].

Personalization is another frontier of IR system development. Approaches such as entity-centric knowledge stores and context-aware prompt augmentation allow retrieval and language models to leverage user history and domain-specific knowledge, improving recommendation quality and contextual relevance [3, 16, 19, 28, 30, 32, 33, 38, 41, 47, 48, 51, 59]. For instance, recent work implements lightweight user-specific knowledge graphs to personalize query suggestion in web and dialogue applications [3, 28].

Despite substantial progress, neural IR models remain vulnerable to adversarial inputs, out-of-distribution queries, and performance degradation under domain shifts [4, 22, 25, 33, 40, 41, 62]. Major contemporary research is focused on: **Robustness**: Addressing adversarial and OOD threats and minimizing performance loss under domain or data distribution shifts [25, 41]. **Interpretability**: Developing tools and evaluation strategies to elucidate neural model decisions and promote responsible deployment [22, 41]. **Benchmarking**: Standardizing evaluation with comprehensive, heterogeneous datasets, such as the BestIR suite [22, 40, 41, 62].

Developing harmonized definitions of robustness and implementing effective defenses for neural retrieval remain critical open challenges, especially as neural retrievers and RAG pipelines are increasingly integrated with LLMs and deployed in knowledge-intensive, real-world domains [22, 40, 41, 62].

2.3 Knowledge and Context Augmentation

Learning Objectives: This section aims to (1) clarify the pivotal goals and techniques of knowledge and context augmentation in LLM workflows, (2) critically analyze the strengths and weaknesses of current augmentation strategies, and (3) situate data-centric methods within the broader context of architectural advances for retrieval-augmented generation (RAG) and knowledge-grounded AI.

Knowledge and context augmentation is fundamental to overcoming domain knowledge gaps and supporting reliable inference in LLMs. In alignment with the overarching survey goals of systematically connecting data-centric and architectural innovation for robust, verifiable AI, this section provides an integrative perspective on state-of-the-art augmentation methodologies.

A range of strategies have been established as key pillars in modern RAG pipelines and data-centric AI. Query expansion and synthetic data generation—utilizing methods such as mixup, chunking, and prompt engineering—address annotation scarcity while increasing diversity and coverage in both model training and inference scenarios [1, 3, 11, 13, 17, 21, 26, 28, 32, 33, 37–39, 47, 51, 53, 55, 60, 63]. Teacher-student knowledge distillation facilitates the transfer of capabilities from large foundation models to smaller, more efficient ones, improving domain adaptation and data efficiency [36, 55, 60]. Active learning and tailored feedback loops support iterative model refinement by leveraging pseudo-labeling and targeted human annotation, which are crucial for annotation efficiency in specialized domains [21, 55, 60]. Notably, chunking and informed context selection, as exemplified by clinical NER pipelines such as CLEAR [3, 33, 37, 39, 42, 51], directly enhance both the accuracy and scalability of information extraction in biomedical NLP.

Integration with knowledge graphs and hybrid neural-symbolic architectures has proven transformative for knowledge-intensive applications. Large language models now interact with, or are augmented by, structured representations such as knowledge graphs to achieve greater factual consistency, verifiability, and support for multi-hop reasoning—crucial for scientific, biomedical, and legal domains [1, 5, 16, 17, 21, 22, 28, 32, 33, 38, 42, 49]. The use of knowledge graph injection enables richer model representations, improving the handling of rare entities and reducing hallucinations, thus aligning LLM outputs with regulatory and safety requirements [5, 22, 33, 42, 49].

A critical challenge common to these augmentation strategies is to balance computational efficiency with responsiveness and the depth of incorporated knowledge [1, 13, 17, 32, 33, 37, 55]. As noted across both biomedical [17, 33, 38, 39, 55, 63] and scientific [3, 5, 21, 28, 49] domains, efficiently selecting and grounding context is central to ensuring factuality while addressing annotation limitations and domain coverage. Modular and hybrid architectures, featuring pluggable augmentation modules, are increasingly prevalent, offering flexibility and supporting explainability, adaptation, privacy, and scalable deployment [21, 36, 55, 60].

As summarized in Table 6, these diverse but complementary techniques underpin advancements in robust, domain-aligned, and verifiable AI built on LLMs. It is important to note, however, that several challenges and open research directions persist: (1) Current methods are limited by the coverage and timeliness of external corpora and knowledge bases [17, 33]; (2) Scalability and computational overhead remain concerns, especially as context lengths and retrieval complexity increase [51, 63]; (3) Handling annotation scarcity, rare or ambiguous entities, and factuality in low-resource domains require further research [13, 21, 37]; and (4) Privacy, fairness, and explainability in data augmentation and hybrid architectures need ongoing attention [17, 33, 36, 60]. Moreover, there is an emerging need for unified taxonomies and frameworks that better integrate data-centric and model-centric augmentation, an area where this survey seeks to motivate further work.

Section Summary: Knowledge and context augmentation strategies—spanning data-centric expansion, active feedback, neural-symbolic integration, and modular design—collectively constitute the backbone of modern retrieval-augmented and verifiable AI. Explicit consideration of their limitations, along with clear linkage to system-level and architectural innovations, is essential for translating research progress into scalable, trustworthy deployments.

2.4 In-Context Data Augmentation Techniques

Within the overarching goals of this survey—namely, to delineate the roles, challenges, and advancements in data augmentation for LLMs and vision models under real-world constraints—this section focuses on recent innovations in in-context data augmentation. As LLMs and vision architectures are increasingly deployed in domains with limited labeled data and rigorous regulatory frameworks, such augmentation strategies have become indispensable to overcome data scarcity and bolster model robustness.

Advanced methods in this space synergize pretrained language models with pointwise information metrics (such as V-information), intent-sensitive filtering, and synthetic data generation. This fusion enhances sample efficiency, particularly for intent detection

Table 6: Representative Knowledge and Context Augmentation Strategies

Strategy	Primary Goal	Exemplar Application Domains
Query Expansion	Increase recall / coverage	Scientific and biomedical IR
Synthetic Data Generation	Address annotation scarcity	Healthcare, vision, surveys
Knowledge Distillation	Efficient adaptation	Low-resource or specialized models
Active Learning / Feedback	Annotation efficiency	Biomedical NLP, legal classification
Knowledge Graph Integration	Factual grounding, multi-hop reasoning	Materials science, clinical, law

and hierarchical text classification tasks [37]. For example, selectively incorporating augmented samples—chosen based on their marginal utility—enables state-of-the-art outcomes while controlling for overfitting and noise, even outperforming naïve in-context prompting approaches in both low- and full-resource scenarios [37].

Parallel innovations have emerged in vision domains. Techniques such as dynamic segmentation and controlled background–foreground composition during data synthesis yield substantial benefits under limited or synthetic data regimes [13]. These results underscore the importance of aligning augmentation strategies with model architectures and the statistical nature of available data.

A salient, practical application is the synthetic augmentation of clinical datasets using open-source LLMs (e.g., LLaMA, Alpaca), deployed locally to preserve data privacy and minimize costs in sensitive health environments where authentic data access is restricted [14]. The integration of carefully curated, high-quality synthetic samples has been empirically shown to robustly improve classifier accuracy under privacy constraints and data scarcity, validating the practical viability of LLM-driven augmentation in clinical and scientific contexts [14].

To summarize within the context of our survey objectives, the evolution of in-context data augmentation encompasses a continuum of strategies:

Intelligent Prompt Engineering: Crafting prompts to generate diverse and contextually relevant synthetic data.

Intent-Aware Sample Selection: Filtering augmented data to retain utility and informativeness.

Domain-Adapted Synthetic Generation: Tailoring generated samples so their statistical and operational properties match the target domain.

The rigorous integration of these strategies into retrieval-augmented models and domain adaptation frameworks is pivotal for building robust, transparent, and high-performing AI systems capable of serving in scientific, clinical, legal, and multimodal environments [1, 3, 11, 13, 14, 17, 21, 26, 28, 32, 33, 37–39, 47, 51, 53, 55, 60, 63].

To maintain cohesiveness as we transition to subsequent topics (including retrieval-augmented generation, contrastive learning, and multimodal augmentation), we will periodically relate each technique back to the survey's central goals: understanding where and how data augmentation most effectively complements broader AI system design.

3 Retrieval-Augmented Generation (RAG) Architectures and Advances

3.1 Core Principles and Process Phases

Retrieval-Augmented Generation (RAG) architectures represent a significant progression in the development of large language models (LLMs), addressing foundational limitations of purely parametric systems—most notably, the prevalence of hallucinations and the constraint of static, outdated knowledge [2, 3, 11, 16, 23, 28, 32, 33, 38–40, 47, 48, 52, 55, 59]. In RAG, the overall workflow is systematically structured into the sequential phases of retrieval, reranking, and generation, forming a tightly-coupled pipeline that enhances reliability across diverse tasks.

The retrieval phase entails the identification of the most pertinent external knowledge sources relative to a user query. This stage encompasses a variety of modalities, such as unstructured texts, structured knowledge graphs, legal documents, and biomedical records [20, 22, 33, 38, 52, 55, 63]. The choice and modernization of retrievers—ranging from traditional sparse-vector approaches (BM25, TF-IDF) to contemporary dense and hybrid models—have proven critical, as these mechanisms determine the informational foundation fed into generative models [2, 32, 33, 38].

Following retrieval, the reranking phase is implemented to reorder candidates by relevance and contextual fidelity. This typically leverages cross-encoder architectures, graph-attention mechanisms, or domain-specific rerankers aimed at optimizing information quality and alignment with user intent [3, 23, 28]. The generation phase synthesizes responses from the curated context using transformer-based decoders, conditioned either on all retrieved evidence or dynamically through focused attention mechanisms [11, 28, 39, 59]. This three-phase procedure has proven to reduce hallucinations, enhance transparency, and ground outputs in verifiable, up-to-date knowledge—impacting clinical, biomedical, and legal domains with demonstrably improved results [40].

RAG's versatility is rooted in the diversity and quality of its underlying knowledge sources:

Biomedical RAG systems incorporate indexed resources like PubMed and UMLS, as well as multimodal clinical records, yielding significant gains in variable extraction and summarization tasks [22, 33, 38, 52, 55].

Legal and regulatory applications ingest multilingual legal texts and case law, enhancing context-awareness and jurisdictional alignment [20, 22, 63].

These heterogeneous sources necessitate advanced strategies—such as data chunking, semantic alignment, and dedicated preprocessing

pipelines—to ensure efficiency and preserve the semantic fidelity of retrieved content [33, 38].

3.2 Architectural Frameworks and Innovations

Progress in RAG systems has evolved from monolithic to modular, interoperable designs that support scalable deployment and sophisticated knowledge integration. High-level RAG data space models (RAG-DSMs) unify the RAG workflow within federated, secure, and interoperable data infrastructures, thereby facilitating cross-institutional knowledge exchange and fostering trust, which is especially crucial in regulated domains [40].

A central advancement in this domain is the emergence of modular retriever-generator pipelines. These architectures not only decouple retrieval and generation modules for greater flexibility but also enable integrated feedback mechanisms, wherein the quality of generated responses can iteratively influence future retrieval phases and vice versa [3, 19, 22, 23, 26, 28, 30, 33, 36, 39, 40, 47–49]. For example, recent work demonstrates the benefits of tightly coupling retrieval and generation both at architectural and training levels, as in Retrieval-Pretrained Transformer models, or by leveraging in-context retrieval augmentation that dynamically improves the factual accuracy of large language model outputs.

Document identifier (docid) management has also seen notable innovation. Approaches such as direct docid generation and generative retrieval models empower systems to support dynamic and scalable retrieval as knowledge resources expand and evolve [33, 38, 61]. This enables more adaptive search, continuous index updating, and faster onboarding of new information without manual intervention.

In addition, cognitive information retrieval (IR) pipelines that blend symbolic reasoning with neural methods have emerged, enhancing interpretability alongside the expressive power of deep learning models [21, 28, 49]. For instance, knowledge graph-augmented pipelines and attention-based subgraph retrieval enable contextually grounded, explainable responses across knowledge-intensive tasks like scientific information extraction and dialogue systems.

A landmark feature across contemporary RAG architectures is their integration with distributed data spaces. These infrastructures support secure data sharing and controlled collaboration among trusted parties within sensitive environments [40]. Such integration underpins organizational interoperability, compliance with regulatory frameworks (e.g., GDPR, HIPAA), real-time knowledge updates, and robust auditing—all while maintaining scalability and low-latency requirements essential for operational deployments.

A high-level comparison of selected architectural innovations is presented in Table 7.

3.3 Advanced Retrieval and Context Management

As RAG models evolve, the sophistication of retrieval strategies has become fundamental to enhancing both performance and adaptability. Recent developments include hybrid retrieval frameworks that synergistically utilize sparse signals (such as BM25) and dense semantic representations, as well as knowledge graph–augmented approaches. These innovations enable more precise and robust information retrieval, outperforming traditional methods in terms of accuracy and contextual relevance, particularly in demanding domains [1, 5, 16, 17, 21, 28, 32, 33, 38, 42, 49]. Advanced techniques, such as attention-driven subgraph construction, allow systems to dynamically select pertinent knowledge subunits guided by the specific context or task at hand. This results in increased retrieval efficiency by narrowing the candidate set to those knowledge elements most likely to support generation accuracy [17, 38, 49]. For instance, attention-based subgraph construction demonstrates improved performance for question-answering over graph-structured knowledge, while entity-augmented retrieval pipelines leverage domain-specific NER to further enhance retrieval focus and minimize irrelevant context [17, 38, 42].

The emergence of logic-of-task (LOT) retrievers and agentic retrieval paradigms—including agentic/LOT-RAG, CRAG, and SRAG—has introduced the ability to configure retrieval procedures dynamically based on user tasks, workflows, and evidentiary demands [33, 40]. Such agent-driven architectures facilitate tight coupling between retrieval and generation, providing critical support for applications requiring transparency, responsiveness, and dynamic augmentation, including clinical question answering and real-time fact verification during health crises [33, 40]. These approaches also support referenced explanations and decrease the risk of hallucination by grounding generated content in verifiable, up-to-date evidence.

Context management, especially for lengthy, unstructured, and highly interdependent documents, remains a significant technical challenge. Context window limitations and the "lost-in-the-middle" effect can lead to vital information being omitted or de-emphasized as input lengths increase [63]. To address this, several strategies have proven effective. Input segmentation divides documents into semantically coherent chunks, improving context retention and reducing the risk of omitting key information [11, 32, 33, 38, 39, 47, 51, 55, 63]. Map-reduce partitioning further scales document processing by parallelizing subdocument analysis and generation, and dynamic context prioritization allows systems to select, prune, or reorder context windows to ensure the highest value information is presented to the model. For example, strategies such as BriefContext partition long retrieval results into concise contexts, mitigating reasoning degradation and improving answer accuracy in medical question answering [63]. Clinical information pipelines utilizing map-reduce-based RAG and advanced context selection approaches, such as entity-driven chunk prioritization, have demonstrated substantial efficiency gains and accuracy improvements in domains like EHR extraction [38, 39, 42, 55].

Best practices are now converging on domain-driven RAG pipelines that treat attributes such as data provenance, security, and transparency as essential metrics, rather than optional considerations [22, 40]. Frameworks like GUIDE-RAG formalize development into structured pre-retrieval, retrieval, and post-retrieval stages, promoting agility in adapting to regulations and continuous system improvement [22, 39]. Transparency, grounded retrieval, and continual evaluation within these iterative cycles are seen as critical to deploying RAG-enabled LLMs responsibly in real-world, high-stakes environments.

In summary, current research trends in RAG emphasize the integration of contextually adaptive, trustworthy, and deeply coupled retrieval-generation systems. Innovations in retrieval methodology, dynamic context management, and pipeline structuring are driving

Table 7: Notable RAG architectural innovations and their domain strengths.

Architecture	Key Innovations	Domain Focus / Strengths
RAG Data Space Models (RAG-DSM)	Federated data access, secure interoperability, regulatory compliance	Clinical, legal, data-sensitive industries
Feedback-Integrated Modular Pipelines	Iterative refinement between retriever and generator; supports adaptive learning	Cross-domain, high scalability
Generative Retrieval	Direct docid generation, dynamic indexing mechanisms	Expanding, evolving knowledge bases
Cognitive IR Pipelines	Symbolic-neural hybridization, enhanced interpretability	Complex reasoning tasks, explainable AI

progress toward scalable, accurate, and transparent deployment of LLMs across the most knowledge-intensive and critical domains.

4 Contextual Data Augmentation, Contrastive Learning, and Multimodal Applications

This section systematically reviews and critically analyzes cutting-edge methods at the intersection of contextual data augmentation, contrastive learning, and multimodal AI applications. The learning objectives for this section are to: (1) clarify foundational paradigms and their influence on state-of-the-art models, (2) enable explicit comparison of methodological strengths, limitations, and trade-offs, and (3) identify outstanding challenges and future research directions. These aims support the overarching survey goal of providing an integrated, nuanced understanding of data-centric and architectural advances in modern AI systems.

We first introduce prevalent approaches to contextual data augmentation, examining how integrating contextual cues influences model generalization, robustness, and transferability. This subtopic directly relates to the survey's objective of exploring data-centric improvements in model performance. Notably, contextual integration can substantially increase diversity and decrease overfitting; however, it may also create domain shift or distort original semantics, restricting applicability in certain domains. A careful analysis of the circumstances in which these limitations arise—such as in real-world multimodal or highly heterogeneous data—remains an open research area. Opportunities exist for more adaptive and domain-sensitive augmentation pipelines that preserve semantic integrity across modalities.

Transitioning to contrastive learning, we discuss the core principles underpinning this self-supervised paradigm, as well as its dominant role in learning invariant and discriminative representations. Emphasis is placed on use cases within both single-modal and cross-modal settings, thereby reinforcing the survey's commitment to bridging data-centric techniques and architectural frameworks. While contrastive learning methods achieve substantial data efficiency and robust feature extraction, they exhibit critical weaknesses: performance is highly sensitive to the choice of positive and negative data pairs, and methods may falter in weakly-labeled or multimodal contexts where pair selection is ambiguous. Deeper investigation into automated selection or generation of informative contrastive pairs—particularly in noisy or dynamic environments—constitutes a promising future direction.

We then examine multimodal applications that integrate data augmentation and contrastive objectives, setting the stage for a unified perspective on how these techniques co-evolve. The synergy between augmentation and contrastive learning has underpinned advances in challenging tasks such as vision-language reasoning, cross-domain retrieval, and adaptive multimodal fusion. Yet, blending these objectives at scale introduces nuanced trade-offs involving data efficiency, computational cost, and maintainability. Further, the integration of context- and modality-aware augmentations with scalable, resource-conscious contrastive learning frameworks remains underexplored.

A critical comparative discussion threads through these subsections, drawing clear connections across data-centric and architectural advances that are at the heart of the survey's organizing framework. To facilitate domain-specific adaptation and future benchmarking, we textualize significant architectural patterns rather than relying exclusively on visual diagrams. The section builds toward a more cohesive narrative by highlighting methodological differences in the integration of contextual cues, augmentation strategies, and contrastive training objectives, enabling practitioners and researchers to align techniques with domain needs.

To concretize open problems and research avenues, we enumerate several actionable directions: developing adaptive, semantically-aware augmentation techniques tailored for multimodal and real-world data; creating robust procedures for constructing or mining high-quality contrastive pairs amid limited supervision; and proposing scalable frameworks that unify augmentation and contrastive paradigms for data- and resource-limited environments. The potential for unified taxonomies that systematically relate contextual augmentation, pair selection, and modality fusion represents a novel framework through which the field may advance.

In summary, this section delivers foundational background and critical insights into the interplay of contextual data augmentation, contrastive learning, and multimodal AI systems. By making the survey's goals explicit throughout the section and providing clear transitions between subtopics, we aim to empower readers with a cohesive, comparative, and forward-looking perspective.

4.1 Contrastive Learning in IR and Recommendation

Contrastive learning has become a foundational approach in modern information retrieval (IR) and recommender systems, enabling the development of richer, more discriminative representations through self-supervised learning objectives. Core frameworks utilize diverse forms of contrast—such as instance-level, multi-view, and augmentation-aware objectives—by forming positive and negative pairs from intrinsic data structures (e.g., user-item interactions, textual co-occurrence) or from synthetic transformations of individual instances. This facilitates robust instance discrimination and enhances representation quality [1, 3, 4, 10, 11, 13, 15, 19, 27, 28, 33, 36, 41, 44, 46–48, 50, 51, 53, 55, 60, 61, 64].

The strategic mining of hard negatives—sample pairs that the model finds challenging to distinguish—serves to refine decision

boundaries. However, imbalance in hard-negative mining may lead to overfitting or instability, necessitating careful tuning of the negative sample selection strategy [11, 28, 48]. Scaling contrastive learning for long-context or sequential data introduces further complexity. Bias towards dominant context patterns can emerge, reducing personalization and diversity in recommendations. Recent works address these limitations by integrating efficient loss functions, hard-negative sampling, and context window mechanisms to preserve scalability while supporting nuanced reranking and mitigating contextual bias [3, 11, 28, 33, 36, 47, 48, 51, 55, 60].

In sequential recommendation, the next-item prediction task has been re-envisioned within a contrastive framework. Models now leverage both context-target and context-context contrast signals to produce contextually sensitive representations. An illustrative example is the ContraRec framework, which unifies these contrastive signals and demonstrates consistent improvements across various sequence encoder architectures and public datasets [54]. This compatibility with mainstream recommendation models highlights the broad applicability of contrastive paradigms.

Building on this foundation, frameworks such as SeqCo further generalize the application of contrastive learning by introducing signals at multiple levels of granularity—including item-wise, batch-wise, and sequence-wise contrast—in sequential recommendation settings. This joint optimization over heterogeneous contrastive losses supports more effective self-supervised representation learning. Empirical results indicate that hierarchical contrast yields superior performance relative to strong baselines, while theoretical analyses reveal the importance of balancing signal intensities and the complexities of instance augmentation [56].

The research emphasis has shifted from merely optimizing encoder architectures towards understanding the synergistic roles of diverse contrastive signals and augmentation strategies in fostering generalizable representations. Hybrid and cross-modal retrieval architectures exemplify this trajectory. These systems frequently integrate multiple modalities—such as text and image—using contrastive loss functions to align semantic information within joint embedding spaces [3–5, 7, 11–13, 17, 19, 21, 27, 28, 30, 33, 36, 44, 46–48, 51, 53, 55, 59–61, 64]. Approaches such as graph-based hashing and deep multimodal transfer learning have been deployed to bridge cross-modal signals, but persistent challenges remain, notably in addressing cross-modal asymmetry (e.g., disparity in information richness between images and text) and label set divergence in domain adaptation. Emerging solutions combine graph convolutional networks with discrete optimization to mitigate these issues, yet quantization loss and sample imbalance present ongoing hurdles [5, 19, 33, 59, 61, 64].

4.2 Contextual Data Augmentation for Neural Models

Objective: This subsection aims to elucidate how contextual data augmentation enhances neural model robustness and generalization across textual, visual, and multimodal applications. Goals are to (1) detail augmentation strategies and their measurable impact (e.g., accuracy improvements, adversarial robustness); (2) analyze limitations and negative outcomes, especially in low-resource or imbalanced data regimes; and (3) synthesize remaining open challenges to guide future research.

Contextual data augmentation is a crucial complement to contrastive learning, as it systematically diversifies the distribution of training instances by manipulating or synthesizing data, thereby supporting increased model robustness and generalization capabilities.

In intent detection, contextual augmentation via prompting large pre-trained language models (PLMs) can synthesize novel utterances. However, inadequate selection and filtering of generated content may introduce semantic drift or noise, sometimes reducing performance rather than improving it. Lin et al. [37] explicitly show that naive in-context prompting does not yield gains for intent detection; instead, careful sample selection, quantified via pointwise V-information (PVI), is necessary to admit only useful augmentations, leading to measured state-of-the-art improvements in few-shot scenarios (e.g., +1.28% in 5-shot and +1.18% in 10-shot settings). This underscores the importance of stringent quality control: excessive or poorly filtered synthetic data can mislead models, cause overfitting to artifacts, or destabilize training, particularly for intents with subtle semantic boundaries.

In the visual domain, simple pixel-level augmentations are often insufficient for industrial or scientific applications featuring imbalanced or limited data. For instance, Kim et al.'s ContextMix [31] addresses industrial defect detection by pasting resized, context-rich regions across batch images, yielding robust, context-aware samples. This yields measurable improvements in classification accuracy, macro F1, and adversarial robustness (e.g., for FGSM and ImageNet-A benchmarks) at minimal computational cost. However, the method's efficacy falls short for extremely small foreground objects, which remain poorly represented in augmented samples—an avenue for future research. Similarly, Dundar and Garcia-Dorado [13] demonstrate that augmenting with foreground-segmented objects and varying backgrounds improves accuracy in low-resource and synthetic data setups, yet caution that inappropriate mixing or poorly defined object/context boundaries can degrade results by confusing the model's inductive biases.

The impact of contextual augmentation is particularly salient in multimodal, multilingual, and personalized tasks, which involve heterogeneous data sources such as text, image, and speech. These scenarios demand versatile augmentation strategies that respect the statistical and semantic properties of each modality. Recent advances in multimodal transfer learning (e.g., cross-modal retrieval and knowledge transfer across disjoint label sets [3–5, 7, 13, 17, 19, 21, 27, 28, 30, 33, 36, 37, 46–48, 51, 53, 55, 61, 64]) show that deep neural models, bolstered with contextually augmented or pseudo-labeled samples, outperform baselines in data-scarce regimes. Nevertheless, common problems persist: semantic misalignment between modalities (e.g., text lacking the objectivity of image data [4]), variability in the quality or relevance of augmentations, and instability in training, especially when intra-class variance is high.

Despite ongoing progress, negative outcomes and open problems are evident. Synthesized or contextually mixed samples can mislead models if the boundaries between objects and context are not well maintained, or if generated samples are of low relevance. Inconsistent augmentation quality may introduce bias, amplify class

imbalance, or destabilize convergence (e.g., see ablation findings in [31, 37]). There is a pressing need for adaptive, assurance-driven augmentation mechanisms—such as dynamic filtering thresholds or modality-aware selection—that address these shortcomings and adapt to the unique needs of each domain.

Key Takeaways and Open Challenges: (1) Many augmentation approaches yield substantial quantitative improvements only when underpinned by rigorous sample selection or semantic checks; careless augmentation can reduce performance or introduce bias. (2) Strategies effective for one modality or domain (e.g., ContextMix for industrial vision) may have clear limitations elsewhere (e.g., tiny object recognition). (3) Consistency, relevance, and semantic alignment of augmentations remain open, measurable objectives. Future work should systematically benchmark adaptive quality control, cross-modal alignment, and negative case analysis to develop robust, generalizable augmentation pipelines.

4.3 Personalization and Adaptive Context

Modern personalization strategies in IR and recommendation critically depend on modeling fine-grained user context, spanning static user attributes as well as dynamic behavioral patterns. Techniques such as user embeddings, adaptive behavioral modeling, and real-time feedback integration facilitate highly individualized information access. Contextual augmentation and contrastive representation learning underpin these user-adaptive systems by enabling models to tailor outputs to users' historical activities and intent filters [3, 37, 38, 55].

Innovative approaches now leverage lightweight entity-centric knowledge representations built from users' search and browsing histories to personalize large language model (LLM) outputs while minimizing privacy risks. Instead of maintaining exhaustive user profiles, these methods project aggregate user interests onto public knowledge graphs, coupling this with session-aware prompt augmentation. The result is improved accuracy and privacy-preserving customization for applications such as query suggestion and open-domain search [3, 38, 55].

Recent work demonstrates that retrieval-augmented generation (RAG) can further enhance personalization by grounding LLM responses in relevant, up-to-date external knowledge sources. Within medical and educational domains, such strategies enable contextually aware, safety-checked, and transparent generation, as shown by frameworks that combine structured user interests, local and external retrieval, and targeted prompt design, leading to notable improvements in relevance, accuracy, and privacy [38, 55].

Data augmentation methods, including selective in-context augmentation with information-theoretic filtering, show promise for intent detection tasks in user-adaptive systems. By generating and selectively incorporating high-value synthetic utterances, models achieve superior performance in low-resource and few-shot settings, highlighting the importance of tailored augmentation to personalization pipelines [37].

However, the transition to real-time adaptation poses significant challenges: managing evolving, non-stationary user preferences; maintaining user privacy and compliance with regulatory frameworks; and scaling adaptive personalization to diverse platforms and linguistic environments.

Key Takeaways and Challenges: (1) Lightweight, privacy-preserving user representations—anchored to public knowledge graphs—offer a scalable alternative to traditional deep user profiling. (2) Retrieval-augmented methodologies and targeted data augmentation improve both personalization fidelity and safety, especially in sensitive domains such as healthcare. (3) Effective adaptive context modeling now requires explicit joint optimization for transparency, fairness, and privacy. Addressing these criteria fuels current interest in federated and on-device learning, privacy-preserving embeddings, and interpretable user modeling frameworks as future research directions.

4.4 Synthesis and Open Challenges

This section synthesizes our survey's explicit objectives—to comprehensively map the landscape of contextual data augmentation and contrastive learning for information retrieval (IR) and recommendation, with measurable goals to (1) chart the taxonomy of current methods, (2) assess comparative strengths and limitations in multimodal, low-resource, and personalization settings, and (3) identify unresolved challenges and emergent research trends. Our literature inclusion methodology, driven by an extensive review of recent works, ensures representative coverage of IR and recommendation studies at the intersection of advanced learning, augmentation strategies, and personalization paradigms.

The joint application of contextual data augmentation and contrastive learning—across both data and model levels—has significantly advanced the ability to meet requirements for modern multimodal and adaptive systems, as well as those operating under data sparsity or personalization constraints. From this synthesis, several noteworthy themes and open challenges have emerged:

Harmonizing Data Augmentation and Adaptive User Modeling: Achieving seamless integration between augmentation strategies and user context remains an unresolved hurdle. The interaction is highly application-specific, with few standard frameworks available. There is an ongoing need for principled methodologies that adapt augmentation dynamically to user profiles and feedback.

Scalability and Adaptability in Contrastive Learning: While contrastive learning shows substantial promise for robust multi-view and cross-modal representation alignment, it continues to face scalability challenges in high-dimensional, sparse, or heterogeneous environments—primarily due to inefficiencies in negative sampling, intensive memory demands, and rigid alignment objectives. Future research should explore adaptive negative mining, memory-efficient architectures, and more flexible contrastive paradigms to overcome these issues.

Ethical, Privacy, and Fairness Considerations: As personalization and context awareness deepen, ethical and privacy challenges intensify. Systematic frameworks are needed to assess and mitigate risks such as bias propagation, data leakage, and inference attacks. Progress toward privacy-preserving augmentation and learning-by-design continues apace, but standardized reporting and evaluation protocols across IR and recommendation domains are lacking.

Evaluation Standardization and Reporting Consistency: A prominent gap is the lack of consensus on evaluation metrics and experimental protocols. We urge the community to develop and

adopt standardized benchmarks for fair and reproducible assessment, especially in the context of cross-modal system generalization and robustness under distributional shifts.

Temporal Progress and Emergent Frameworks: Foundational studies laid the groundwork by demonstrating the utility of augmentation and contrastive techniques, while the past two years have seen the emergence of frameworks that jointly optimize augmentation procedures and contrastive objectives, increasingly emphasizing privacy and fairness from the outset. These developments mark a shift toward holistic system design and highlight the growing maturity of the field.

Key Takeaways:

1. **Integration Complexity:** The synergy between augmentation and personalization needs further methodological clarity.
2. **Scalability:** Computational bottlenecks in contrastive learning remain a central concern for large-scale heterogeneous data.
3. **Responsible AI:** Ethical imperatives and privacy-by-design principles are crucial as personalization deepens.
4. **Standardization:** Community efforts toward unified evaluation and reporting will boost comparability and reproducibility.
5. **Emerging Trends:** Recent frameworks emphasize joint optimization and responsible, domain-transferable system design.

Ongoing advances in augmentation strategies, cross-modal alignment mechanisms, and privacy-centric modeling are vital for developing IR and recommendation systems that are robust, fair, and scalable, ensuring readiness for future demands and interdisciplinary applications. We encourage the adoption of standardized evaluation practices and the continued exploration of holistic, privacy-aware, and adaptive models to address the remaining open challenges in the field.

5 Applications in Biomedical, Legal, and Cross-Domain Contexts

This section aims to provide an explicit and structured survey of AI applications across the biomedical, legal, and cross-domain fields, clarifying both the measurable objectives and the inherent limitations of current approaches. Our primary objectives here are to: (1) systematically review key methodologies and achievements in each domain, outlining both successes and failed attempts/negative results where documented; (2) elucidate domain-specific challenges, with particular emphasis on unresolved issues, limitations, and negative findings; and (3) synthesize emerging cross-domain strategies, highlighting integrative innovations that leverage interdisciplinary insights. These goals are intended to give readers a clear framework for interpreting the breadth and depth of AI's role in these critical sectors.

To maximize clarity and accessibility, the section is divided into dedicated subsections focusing separately on biomedical, legal, and cross-domain contexts. Each subsection describes representative advancements, mainstream techniques, and unique integration efforts, while explicitly critiquing methodological limitations and the rationale behind any omitted or lightly-treated subtopics. Wherever appropriate, key findings, challenges, and domain-specific insights are visually separated or summarized to facilitate standalone accessibility and reference.

In recognition of the evolving research landscape, we conclude this section with a synthesizing table that encapsulates open research problems and persistent challenges spanning all three domains, offering a concise reference for future investigation and highlighting opportunities for cross-disciplinary learning. This comprehensive approach emphasizes not only the achievements but also the constraints and obstacles shaping AI's ongoing impact in biomedical, legal, and cross-domain scenarios.

5.1 Clinical and Health Applications

The integration of Retrieval-Augmented Generation (RAG) into large language model (LLM) pipelines has produced transformative advances within the clinical landscape, addressing core limitations of LLMs such as hallucinations, temporal staleness, and opaqueness in decision provenance [5, 16, 17, 21, 24, 29, 33, 35, 38, 39, 42, 43, 46, 52, 55, 62]. In clinical question answering and decision support, RAG-enabled systems routinely surpass unaugmented LLMs in accuracy by systematically grounding outputs in current, domain-specific guidelines and contextual patient data. For example, SurgeryLLM—a domain-adapted RAG framework—demonstrated improved performance across all core clinical tasks, including lab value interpretation and operative note generation, by directly aligning recommendations to national standards and reducing uncertainty or outright refusal evident in baseline LLM outputs [43].

Comparative benchmarking has consistently shown state-of-the-art RAG architectures, especially those leveraging international guideline corpora alongside advanced retrievers and models such as GPT-4, can exceed expert clinician accuracy in perioperative scenarios. These systems also improve reproducibility and safety, while significantly minimizing workflow inconsistencies and potential surgery cancellations [29].

Infrastructure-level enhancements have been realized through RAG integration into electronic health records (EHRs), exemplified by the CLEAR pipeline. CLEAR combines clinical named entity recognition with RAG-based chunk retrieval, enabling near-real-time extraction of structured variables from narrative notes with far fewer computational resources compared to dense embedding-based approaches. This preserves contextual integrity, avoids degradation commonly observed in long-context LLMs, and facilitates scalable, automated construction of clinical knowledge graphs for downstream applications [42]. Moreover, multi-task frameworks like RAMIE operationalize RAG via task-specific prompting and simultaneous learning, yielding substantial gains in extracting complex dietary supplement information and further demonstrating RAG's flexibility and efficiency when paired with targeted retrieval mechanisms [16].

Beyond structured decision support, RAG has proven vital in constructing biomedical knowledge bases, literature recommendation engines, and patient-facing educational tools. Systems such as RefAI synthesize and summarize literature with traceable citations, thereby fundamentally reducing hallucinations and data fabrication commonly observed in prior LLM pipelines. This is achieved by coupling retrieval from validated sources (for example,

Table 8: Summary of Open Research Problems Across Biomedical, Legal, and Cross-Domain AI Applications

Domain	Key Open Problems	Challenges Highlighted	Notable Negative Results/Limitations
Biomedical	Generalizability to heterogeneous data; Explainability; Regulatory compliance	Scarcity of labeled data; Privacy-preserving learning; Bias mitigation	Limited reproducibility in clinical settings; Failure to translate from theory to practice
Legal	Interpretability of legal reasoning; Handling evolving legislation	Complex, ambiguous data; Lack of standardized datasets; Domain adaptation difficulties	Low accuracy on rare case types; Overfitting to historical biases
Cross-Domain	Model transferability; Unified representation learning	Integration of disparate domain ontologies; Scalability across contexts	Negative transfer effects; Semantic drift across domains

PubMed) with advanced summarization capabilities [17, 62]. In addition, RAG-enabled knowledge graph augmentation is now central to automated biomedical knowledge synthesis, leveraging LLMs for both extraction and semantic structuring of vast, heterogeneous literature, which in turn advances chain-of-thought reasoning and accessibility for clinicians and researchers [21, 38, 39].

A prevailing research focus centers on factuality and safety, especially for deployments sensitive to misinformation and fact-checking, such as in public health (e.g., infodemic detection during the COVID-19 pandemic). RAG-augmented LLMs—particularly those employing agentic deliberation or layered retrieval—outperform standard LLMs at identifying and contextualizing misinformation. These models provide transparent, referenced justifications, thereby enhancing user trust and actively countering automation bias [2, 33, 52, 63]. The introduction of factuality modules, stance rerankers, and document-driven generation has significantly increased the accuracy and explainability of health information retrieval, as documented by measurable improvements in established benchmarks [33].

RAG and LLM pipelines have also accelerated social media and public health analytics by supporting disease trend detection, transfer learning for emergent events, and annotation benchmarking [20, 29, 49, 50, 57]. Adaptive retrieval and summarization, particularly through zero- and few-shot transfer, enhance model agility in rapidly evolving domains and in low-resource settings, thereby facilitating early warning and rapid response to emerging health threats [29, 37, 49, 50, 55, 57].

Nevertheless, persistent challenges remain. Qualitative research highlights that, while NLP approaches are efficient for thematic extraction from survey data, they continue to lack the interpretive depth and contextual sensitivity of expert human qualitative analysis, particularly when processing slang or subcultural language [18]. As such, hybrid analytic frameworks that combine rapid NLP-based analysis with human interpretive oversight consistently yield superior insights. More broadly, RAG architectures—although effectively mitigating issues of factuality and recency—are ultimately limited by the scope, quality, and update latency inherent in their external knowledge sources [37, 55, 63]. Continued research is addressing the refinement of context-aware retrieval granularity, dynamic knowledge updating, and bias mitigation, alongside infrastructure and privacy constraints relevant to real-world clinical deployment [5, 17, 21, 24, 33, 38, 46, 52, 55].

Section Takeaways and Key Challenges:

The integration of RAG into clinical and health domains has yielded substantial improvements in accuracy, factuality, and workflow efficiency. However, measurable and ongoing challenges remain, including:

– **Dependence on external knowledge source quality and update latency**: The accuracy and reliability of RAG-augmented systems are fundamentally constrained by the recency, coverage, and curation quality of the underlying corpora [33, 37, 55, 63].

– **Contextual gaps in processing long or unstructured data**: Effective EHR and clinical note processing requires continued innovation in context-aware retrieval to prevent loss of information or interpretive nuance [42, 63].

– **Interpretability and oversight**: Despite increased transparency compared to baseline LLMs, RAG systems may still generate outputs that lack interpretive depth, particularly in complex qualitative analyses, reinforcing the importance of integrating human oversight [18].

– **Privacy and infrastructure**: Real-world deployment in healthcare is limited by data privacy constraints, pipeline scalability, and system integration challenges [5, 21, 38, 46].

In summary, while RAG pipelines have markedly improved clinical, biomedical, and public health NLP applications, overcoming data quality, update frequency, interpretability, and privacy barriers is a measurable priority for future research and deployment.

5.2 Legal, Regulatory, and Security Applications

In legal and regulatory contexts, RAG-based pipelines must simultaneously deliver advanced functionality such as complex question answering, document analysis, and compliance support, while rigorously adhering to sectoral requirements for security, explainability, and operational trustworthiness [22, 40]. Recent legal pipeline architectures increasingly employ retrieval-augmented systems to facilitate transparent decision-making, robust cross-referencing of statutes and precedent, and demonstrable provenance—characteristics vital for high-stakes legal reasoning [22]. Integration of secure, interoperable RAG frameworks within legal and healthcare infrastructures further supports the acute demands for privacy, auditability, and risk containment. These demands are reinforced by a maturing standards landscape that prioritizes transparent and well-documented pipeline operations [22, 40].

Privacy-preserving data architectures are especially emphasized. Compliant retrieval mechanisms—including federated and decentralized data handling—help to ensure that sensitive client or patient information stays protected throughout the RAG pipeline [3, 7, 8, 10, 19, 22, 25–28, 33, 34, 36, 41, 45, 50, 51, 53, 55, 60, 61, 63]. Although advancements in privacy and compliance are notable, critical limitations include persistent trade-offs between retrieval efficiency and the risk of privacy leakage, notably in cross-jurisdictional and multi-tenant deployments. Additionally, adversarial and out-of-distribution risks remain significant for neural IR systems [41], necessitating ongoing improvements in both detection and robustness strategies within RAG frameworks.

Recent research foregrounds the imperative for rigorous risk management alongside practical functionality. Integrated solutions now include risk-aware retrieval strategies, policy-constrained generation modules, and traceable attribution of knowledge sources to withstand adversarial scrutiny and legal discovery requirements [3–5, 7–11, 15, 16, 19, 22, 25–29, 32–34, 36, 39, 41, 42, 45–48, 50, 51,

Table 9: Summary of Key Benefits and Ongoing Challenges of RAG in Clinical Applications

Application Area	Key Benefits	Ongoing Challenges
Clinical Q&A & Decision Support	Grounding in current clinical guidelines Increased accuracy and safety	Reduced workflow inconsistencies Dependence on external source quality Update latency
EHR Data Extraction	Real-time structured variable extraction Resource efficiency	Scalable knowledge graph construction Context loss in long/unstructured notes Privacy management
Biomedical Knowledge Synthesis	Factually grounded literature summarization Traceable citations	Hallucination in absence of relevant sources Information overload
Public Health Analytics	Early detection of disease trends Enhanced model agility via zero-/few-shot transfer	Data sparsity in emerging domains Sustained need for human oversight

53, 55, 59–61, 63, 64]. However, practical deployments reveal that maintaining explainability and transparency often comes at the expense of system efficiency, particularly in complex multi-step pipelines.

A significant requirement in legal decision support is explainability. Legal professionals require not only accurate answers but also actionable rationales anchored in statutory law, caselaw, and procedural precedents. Retrieval-augmented systems enable traceable chains of reasoning and counterfactual analysis, supporting a solid foundation for future explainable legal AI systems that can satisfy both regulatory and societal expectations [22]. Hybrid systems that combine retrieval-augmented generation with formal logic and argumentation models have been proposed to bridge the gap between output fluency and transparency, which is essential for increasing interpretability in high-stakes legal settings [9]. Nonetheless, current systems face limitations in the scalability of argument mining, integration with LLMs, and open-domain applicability [9].

Despite the advances described, multiple open research challenges remain and must be explicit focal points for future development. The following outline the most pressing issues, along with desired measurable outcomes where possible:

Cross-jurisdictional Scalability: Development of RAG pipelines that are empirically validated to operate across multi-jurisdictional and cross-lingual legal scenarios, with performance assessed on legal QA and document review tasks covering diverse regional statutes.

Transparency vs. Efficiency: Explicit benchmarking of workflow transparency (e.g., frequency and clarity of source citation, policy traceability rates) against efficiency measures (e.g., response latency, computational overhead) in legal practice deployments.

Explainability: Empirical improvements in output interpretability and auditability, quantifiable by rates of citation accuracy and reason-chain completeness, through the integration of argumentation engines or structured reasoning frameworks [9].

Negative Cases and Limitations: Known persistent challenges include contextual augmentation failures (e.g., retrieval irrelevance or fact misalignment), privacy-preserving IR vulnerabilities to inference attacks or data leakages, and robustness to adversarial input or out-of-distribution scenarios [41]. Quantitative evaluation on privacy risk and retrieval robustness benchmarks should be standard practice in this domain.

Key Takeaways: RAG-based legal and regulatory systems demand measurable advances in privacy preservation, transparency, and argument-driven explainability, but must also address lingering gaps in cross-jurisdictional adaptability and adversarial robustness. The most impactful contributions will concretely demonstrate improvements using standard benchmarks for legal, privacy, and security tasks, as well as transparent reporting of negative results and system limitations.

Summary of Measurable Objectives:

Goal	Measurable Outcome
Privacy Preservation	Differential privacy/precision-recall (on legal IR
Explainability	Rate of rationale citation/completeness
Transparency	Legal provenance trace rate
Cross-jurisdictional Scalability	Accuracy on multi-lingual, multi-region data
Robustness	OOD/adversarial error rates

By clearly articulating and consistently evaluating these explicit objectives, the field will be better equipped to iteratively improve and rigorously assess legal, regulatory, and security-focused RAG pipelines.

5.3 Vision and Multimodal Cross-Domain Applications

The principles underpinning RAG have been extended beyond text, with recent studies successfully applying retrieval-augmented pipelines to vision and multimodal knowledge enrichment. This expansion has significant ramifications across scientific, technical, and operational domains [3–5, 7, 8, 13, 17, 19, 21, 27, 28, 30, 33, 36–39, 42, 46–48, 50, 51, 53, 55, 61]. In the context of visual recognition, techniques such as foreground/background separation and

synthetic data generation have improved object classification performance—particularly in data-constrained or specialized scenarios. Notably, when augmentations such as context background manipulation and object segmentation are applied, classification accuracy in convolutional networks is enhanced, especially for limited-data and synthetic datasets [13]. When these augmentations are incorporated into multimodal RAG architectures, they enrich contextual retrieval for downstream tasks by providing diverse, information-rich representations [13].

Modern pipelines increasingly facilitate multimodal and cross-lingual retrieval, enabling the integration and joint reasoning across not just text but also images, graphs, and tabular data. Key enabling technologies include deep multimodal transfer learning, cross-modal hashing with graph convolutional networks to address semantic feature alignment and information asymmetry between modalities [4], and the deployment of optimized index/search strategies for retrieval in complex scientific and legal domains where exhaustive labeled data are scarce [13, 37, 47, 48]. For example, neural architectures that generate enriched image descriptors by combining significant convolutional activations with fully connected layers have produced retrievals that closely match query images not just semantically, but also in visual properties such as background, texture, and color distribution [50].

These technical advances are particularly valuable in domains where evidence spans various modalities—such as documents, figures, structured databases, and knowledge graphs—supporting enhanced vision-language models for document analysis, benchmarking, and collaborative scientific workflows [3, 5, 7, 8, 17, 19, 21, 28, 30, 33, 36, 39, 46, 48, 51, 55, 61]. For instance, in scientific and biomedical applications, multimodal RAG pipelines have demonstrated improved factual accuracy, reduced hallucination rates, and increased transparency in downstream tasks by integrating large language models with specialized retrieval modules over both structured and unstructured knowledge sources [5, 17, 33, 38, 39, 55].

As these trends accelerate, the move towards scalable, multimodal RAG systems highlights the central challenge of trustworthy and efficient knowledge integration within mission-critical environments. Regardless of deployment context—be it biomedical, legal, or scientific—the most effective RAG pipelines are those which expand accessible knowledge while upholding rigorous standards of explainability, privacy, and domain adaptability.

6 Benchmarking, Evaluation, Security, and Interpretability

This section provides an integrated and quantitatively focused overview of the methodologies and challenges in benchmarking, evaluation, security, and interpretability of AI models. The section's objectives are (1) to systematically review benchmarking practices for AI systems by categorizing benchmark types and delineating their quantitative criteria; (2) to enumerate and analyze evaluation metrics and approaches, explicitly highlighting diverse assessment criteria and their comparative strengths; (3) to synthesize major security vulnerabilities and the most prominent safeguards, noting distinct threat models without redundant repetition; and (4) to map interpretability techniques to their practical implications, with a particular focus on frameworks fostering explainability and trust.

These focal topics represent essential and interconnected components in the AI development and deployment lifecycle: Benchmarking establishes reproducible and fair grounds for model comparison across standardized datasets, with this survey contrasting task-specific, general, and adversarial benchmarks along dimensions such as size, diversity, and application scope. Evaluation approaches are outlined with precise quantitative metrics—such as accuracy, F1-score, AUC for classification; BLEU, ROUGE for NLP; and robustness against adversarial perturbations—each sub-section clearly numbered and distinctly headed for ease of reference. Security analysis summarizes vulnerabilities and countermeasures, cross-referencing privacy preservation and attack resistance, while minimizing recurring discussions by directly integrating points about data, models, and environments. Interpretability subsection details taxonomies of explainability techniques, contrasting model-intrinsic, surrogate, and user-centered frameworks, and underscoring their alignment with ethical and regulatory objectives.

Transitions between these core areas are explicitly strengthened as follows: We first examine benchmarking (Section 4.1), then advance to evaluation criteria (Section 4.2), address security considerations (Section 4.3), and conclude with interpretability (Section 4.4), culminating in a synthesis that highlights how effective benchmarking and evaluation facilitate robust security frameworks and interpretable AI deployment. These section-specific goals are explicitly cross-referenced to the survey-wide objectives from the Introduction to ensure seamless integration and cohesion.

The upcoming structured subsections are as follows: Section 4.1: Benchmarking Methodologies and Dataset Taxonomies Section 4.2: Evaluation Metrics and Comparative Criteria Section 4.3: Security Threats and Defense Mechanisms Section 4.4: Interpretability Techniques and Frameworks

The section concludes with a unified summary that distills novel integration insights, summarizes highlighted taxonomies and frameworks, and formulates open challenges and promising research opportunities tightly mapped to the survey's overarching aims.

6.1 Evaluation Protocols and Standards

Rigorous evaluation is a foundational requirement for the deployment of retrieval-augmented generation (RAG) and large language model (LLM) systems, especially in domains characterized by high stakes, regulatory oversight, and complex data modalities. Contemporary evaluation frameworks extend well beyond traditional accuracy metrics, embracing a nuanced matrix of criteria—including robustness, factuality, explainability, personalization, and data quality—that reflect the diverse requirements of stakeholders and deployment scenarios [3, 5, 7, 8, 10, 13, 16, 19, 20, 24–26, 28–30, 32–34, 36–39, 41, 42, 45, 46, 49–52, 55, 60, 63].

While accuracy remains the most extensively reported metric, it alone is insufficient to capture the multi-dimensional nature of real-world RAG and LLM performance. Robustness, measured by a system's resilience to distributional shifts and adversarial perturbations, is critical—particularly in open or adversarial environments. The limitations of pointwise evaluation have become clear as recent robust information retrieval (IR) benchmarks have

demonstrated the necessity of systematic adversarial and out-of-distribution (OOD) testing in addition to innovations in model architecture [33, 37, 45, 63].

Factuality presents a persistent challenge: although RAG systems aim to mitigate the hallucinations typical of parametric models by grounding responses in verifiable external sources, ensuring both the veracity of cited content and its correct alignment with generated answers remains an unresolved methodological hurdle [3, 13, 20, 22, 24, 28, 33, 34, 36–38, 40, 55, 60, 63].

Explainability and interpretability have risen to equal importance alongside accuracy, driven by regulatory mandates and the growing demand for model transparency. Evaluation now incorporates both mechanistic interpretability—diagnosing internal logic and causal pathways in deep architectures—and model-agnostic techniques, such as output rationalization, feature attribution, and counterfactual simulation [3, 6–8, 17, 22, 24, 33, 34, 36, 38, 40, 46, 55, 60]. An increased emphasis on user- and context-centered evaluation, particularly for clinical and scientific risk audits, has prompted the widespread adoption of human-in-the-loop benchmarks and mixed-method studies, combining quantitative metrics with expert qualitative assessment [5, 7, 10, 16, 24, 26, 32, 33, 44].

Personalization has emerged as a critical standard as RAG/LLM-based systems are increasingly tailored to reflect individual user histories, preferences, and knowledge profiles, all while maintaining privacy and scalability [13, 19, 20, 33, 37, 38, 52, 55]. Notable advances, such as entity-centric knowledge projection and context-augmented prompting, have demonstrated substantive gains in system relevance and user satisfaction, particularly in applications such as web and health information retrieval [20, 55].

A key innovation in data-centric evaluation is the use of information-theoretic sample filtering, including pointwise V-information (PVI). Such approaches enable the quantification and curation of valuable training samples, reducing dataset redundancy and noise, thereby leading to improved model generalization and performance—especially in few-shot and low-resource contexts [13, 24, 37]. Ablation studies also remain essential for disentangling the contributions of individual architectural or data-driven components, facilitating reproducible synthesis across various modalities and thematic domains [3, 13, 20, 22, 24, 28, 33, 34, 36–38, 40, 55, 60, 63].

As detailed in Table 10, effective evaluation of RAG and LLM-driven systems demands a multi-faceted approach that integrates these considerations to address real-world complexities.

6.2 Benchmarks and Datasets

Benchmarking RAG and LLM systems requires access to diverse, high-quality datasets that are representative of relevant tasks and domains. In biomedical and clinical research, established benchmarks such as PubMed, MIMIC, UMLS, BioASQ, MedQA-US, and MedMCQA allow for rigorous evaluation of knowledge-intensive and reasoning tasks, while systematic frameworks like GUIDE-RAG offer structured stages for clinical RAG implementation [1–5, 7–11, 13, 15, 16, 20, 21, 25–30, 32–39, 41, 42, 46–50, 52, 53, 55, 57–60, 62–64]. For social media and open-domain conversational applications, Twitter datasets and OpenDialKG are widely used to evaluate LLM-based systems in highly dynamic, less-structured environments.

The growing use of synthetic datasets supports robust evaluation, particularly for continual compositional inference and adversarial out-of-distribution (OOD) robustness testing [13, 33, 37, 55]. However, each dataset category presents unique advantages and limitations. Annotated benchmarks in domains such as clinical or legal settings provide structured, interpretable evaluation but are constrained by scalability, the need for continual updates to maintain gold standards, and potential coverage bias. Open-domain and vision-oriented datasets, such as IMAGENET1M and MatSci, facilitate broader generalization assessment but may lack the detailed annotation necessary for fine-grained reasoning. A notable gap remains in the availability of well-annotated multilingual and multimodal datasets, which limits advancements in cross-lingual and cross-domain generalization.

Recent progress in knowledge graph extraction and domain adaptation, exemplified by resources like MatSciBERT and KG-FM in materials science, as well as advances in multimodal benchmark synthesis, have expanded evaluation capabilities beyond text-only tasks [2–4, 28, 36, 37, 49, 58, 62, 64]. Nonetheless, the persistent shortage of benchmarks featuring naturalistic, user-generated queries paired with gold-standard annotations—especially in non-English languages—continues to impede comprehensive end-to-end evaluation. Addressing this critical gap will require collaborative dataset curation efforts and standardization initiatives to increase benchmarking rigor, diversity, and the practical assessment of RAG/LLM systems.

6.3 Interpretability, Security, and Human-in-the-Loop

Interpretability, security, and human oversight are increasingly vital dimensions in the evaluation and deployment of RAG systems as they transition into mission-critical and societally impactful domains. This section threads practical engineering challenges and analytic considerations relevant to these aspects, highlighting their interplay with the overall survey objectives: to assess how RAG advances trustworthiness, reliability, and real-world applicability in diverse settings beyond high-stakes clinical and legal contexts.

Evaluation strategies are evolving toward user- and context-centered risk audits, emphasizing transparency and causal traceability of outputs—imperatives in domains ranging from healthcare [3, 5–10, 16, 17, 22, 24, 26, 27, 29, 32–34, 36, 38–42, 44, 46, 55, 59, 60] and science [5, 46] to open-domain information access [6–8]. Explainability requirements now extend beyond retrospective justifications, demanding prospective rationales that enhance user trust, facilitate troubleshooting, and support regulatory compliance [3, 6–8, 22, 33, 34, 36, 38, 40, 55, 60]. Causal interpretability frameworks, including those that attribute predictions or errors to specific model components or data features, enable targeted debugging and continual improvement—for example, through mechanistic analyses in neural IR systems [7, 17, 22, 33, 40, 44, 46, 55]. Despite these advancements, persistent limitations include model opacity, context truncation, handling of ambiguous or contradicting information, and integration with user workflows [6–8, 20, 24, 32, 33, 39, 42, 45, 55].

Comparative evaluation protocols—combining human and LLM-based annotation—facilitate large-scale benchmarking but reinforce the necessity for domain experts in adjudicating subjective or

Table 10: Principal Evaluation Criteria and Representative Methods/Frameworks in RAG/LLM Assessment

Evaluation Criterion	Description	Representative Frameworks / Considerations
Accuracy	Overall correctness of model outputs on benchmark tasks	Standard performance metrics (e.g., exact match, F1), task-specific scoring
Robustness	Resilience to distributional shifts, adversarial inputs, or OOD data	Adversarial/OOD testing protocols, stress-test suites
Factuality	Faithfulness of outputs to external knowledge or ground truth	Source attribution, hallucination detection, citation alignment metrics
Explainability/Interpretability	Transparency and causal traceability of model predictions	Mechanistic analyses, rationalization, feature attribution, counterfactual studies
Personalization	Adaptation to individual user context, preferences, or history	Contextual retrieval, entity-aware prompting, privacy-preserving personalization methods
Data Quality/Curation	Value, diversity, and relevance of datasets used for training and evaluation	Information-theoretic filtering (e.g., PVI), annotation standards, ablation studies

context-dependent outputs [6–8, 20, 24, 33, 45, 55]. Human-in-the-loop designs are especially critical in domains such as scientific discovery [46], clinical recommendation [24, 29, 39, 55], legal technology [22], and personalized recommendation [3, 34, 36, 60], ensuring contextual scrutiny and calibration of user trust. Notably, studies in areas such as document retrieval and information management show that user-involved organizational practices and transparent model logic significantly enhance both retrieval efficiency and perceived system reliability [6–8].

Security and privacy also pose engineering and deployment challenges as RAG is adopted across healthcare, legal, and increasingly open or federated data ecosystems. Key imperatives include privacy-preserving computation, trustworthy data sharing, and regulatory alignment, motivating innovations such as RAG integration with secure data spaces, federated learning, and granular access controls [22, 40]. Striking a balance between data utility and privacy—particularly across institutional or jurisdictional boundaries—remains an open technical and legal challenge [22, 40]. Frameworks such as RAG4DS [40] and privacy-aware RAG for recommender systems [3, 36, 60] outline emerging patterns but highlight that standardized solutions are nascent.

To aid synthesis, we summarize representative evaluation results and benchmark datasets that address these challenges (see Table 11 below):

Security and robustness remain cross-cutting concerns [22, 40, 41], with adversarial and out-of-distribution vulnerabilities, privacy threats, and legal ambiguity constituting ongoing debates. Recent surveys [22, 40, 41] stress the lack of harmonized robustness benchmarks and universal defense mechanisms, calling for research into OOD generalization, continual adaptation, and policy-compliant engineering.

Open Research Problems and Future Directions:

A synthesis of the reviewed literature highlights several persistent gaps and future research priorities in RAG interpretability, security, and human involvement:

Integrative Summary:

Interpretability, security, and robust evaluation in RAG systems are intricately linked with both analytic goals and practical deployment. As elucidated above, the path toward trustworthy, effective AI requires coordinated advances across technical methodologies, user-involved design, and comprehensive regulatory frameworks. Persistent open questions—including the standardization of evaluative and privacy protocols, scalable human-in-the-loop deployment, and the construction of robust benchmarks—underscore that solutions will demand ongoing interdisciplinary collaboration at the intersection of technical innovation, human factors, and policy expertise.

7 Robustness, Ethics, Responsible Deployment, and Workflow Integration

This section examines the interplay between robustness, ethical considerations, responsible deployment, and workflow integration in the context of Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) systems. The objectives here are to: (i) provide a clear account of the technical and socio-technical challenges in these domains; (ii) explicitly link these aspects back to the overall survey goals of promoting reliable, transparent, and societally aligned AI deployment; and (iii) synthesize open problems and future directions for each area.

7.1 Robustness

Section Objective: This section aims to clarify the evolving challenges of achieving and measuring robustness in Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) systems, and to distinguish the unique integrative perspective offered by this survey compared to prior reviews.

Robustness is a cornerstone of trustworthy RAG and LLM systems, ensuring consistent performance under a range of input distributions and adversarial scenarios. While previous surveys have catalogued individual threats and classical defenses, our synthesis emphasizes a cross-domain perspective: how robustness concerns and mitigation techniques interconnect across varying deployment settings, highlighting gaps in field-wide versus domain-specific approaches.

Contemporary evaluation indicates that models remain susceptible to prompt injection, distributional shift, retrieval errors, and adversarial attacks. Current mitigation strategies—including adversarial training, ensemble retrieval methods, and fallback mechanisms—have achieved partial success but also reveal limitations in generalization across domains and adversarial resistance. For example, while adversarial training may yield improvements in general NLP applications, its domain transferability remains limited in specialized settings. This survey contrasts these patterns, drawing attention to the differential impact of robustness strategies in field-wide versus vertical-specific environments.

Notably, the gap in continuously monitoring robustness and systematically stress-testing deployed workflows represents an area of ongoing debate. Failed approaches, such as overreliance on static benchmarks or excessive regularization that undermines utility, underscore the importance of flexible robustness evaluation frameworks. Consider the cautionary example of deploying static evaluation pipelines in proactive customer support bots versus healthcare diagnostics—these highlight unique challenges requiring tailored robustness checks.

Table 11: Sample evaluation results highlighting interpretability, reliability, and human-in-the-loop challenges in RAG systems across domains

Domain	Task	RAG System/Framework	Notable Results	Limitations/Challenges
Clinical NLP	Variable Extraction	CLEAR [42]	Avg F1: 0.90–0.97; Inference time: 1.04–4.95s/note	Focus on structured data; further deployment needed
Medicine	Fact-Checking	Self-RAG [33]	Accuracy: 0.973; Referenced explanations	Corpus coverage; dependency on reference data quality
Oncology/Clinical Trials	Recommendation	Retrieval-aug. GPT-4 [24]	Precision: 63%; Recall: 100%; F1: 0.77	Modest sample size; single-center study
Diabetes Education	Patient QA	RISE [55]	Accuracy gain: 7% (G-4); Comprehensiveness +0.44	Scope limited to selected domains/queries
Information Management	Personal Retrieval	Active storage [7, 8]	Mistake/failure reduced to 3–15%; Retrieval 34s–87s	Cognitive burden; domain generalizability
Legal Tech/Data Spaces	Secure RAG	RAG4DS [40]	Unified lifecycle and privacy framework proposed	Standardization; practical deployment

Table 12: Summary of Open Research Problems and Future Directions in RAG Interpretability and Security

Challenge	Open Problems and Future Directions
Interpretability	Causal traceability in complex pipelines; transparent rationale generation for users and regulators; evaluation of proactive vs. post hoc explanations; bridging model, data, and workflow transparency [6–8, 33, 36, 44, 46, 55, 60]
Security and Privacy	Privacy-preserving computation and federated data sharing; harmonized standards for regulatory compliance; adversarial/robustness benchmarks; transparent access controls; deployment in federated, cross-jurisdictional environments [22, 36, 40, 41, 60]
Human-in-the-Loop Integration	Effective adjudication frameworks; workflow-aligned user interfaces; scalable hybrid evaluation; leveraging user expertise across diverse domains (e.g., clinical, scientific, open-domain) [6–9, 16, 22, 24, 33, 39, 40, 46, 55]
Benchmarking and Evaluation	Unified, multi-domain benchmarks for risk, interpretability, and OOD robustness; efficient data augmentation using LLMs; standardized protocols for subjective tasks and human+LLM assessment [20, 22, 33, 36, 38, 40, 41, 45, 55, 60]
Practical Engineering Constraints	Managing computational costs, latency, and scaling; integrating RAG architectures into existing IT infrastructure, including EHR and data spaces; modularity for generalization and ongoing adaptation [3, 5, 22, 32, 36, 39, 40, 42, 60]

Integrative Summary: Robustness underpins nearly all responsible deployment objectives. Yet, existing strategies require deeper integration with monitoring and feedback in production workflows. This survey uniquely draws together analytic advances and practical engineering solutions, proposing a roadmap where system-level robustness adapts to both evolving threats and domain-specific exigencies. The open questions in robustness, as synthesized here, are thus twofold: which techniques generalize across all RAG/LLM deployments, and which must be highly tailored to sensitive use cases? Advancing on these fronts remains critical for the responsible future of LLM-powered systems.

7.2 Ethics and Responsible Deployment

Ethical challenges in RAG and LLM deployment include bias propagation, privacy violations, opacity in output provenance, and disparate impact across user groups. Recent frameworks stress the importance of pre-deployment audits, transparency mechanisms, active user consent, and explicit risk assessments. However, the translation of theoretical principles into enforceable practice remains contested. In particular, attempts to fully automate ethical compliance have highlighted difficulties in operationalizing nuanced human values and adapting to novel contexts.

Limitations persist in evaluating bias and harm across emerging application domains, especially beyond well-studied clinical or legal environments. Ongoing debates focus on the balance between model autonomy and human oversight, and the scalability of post-hoc ethical interventions. Standardizing best practices in risk documentation and user-facing disclosure emerges as a prominent open problem.

Integrative Summary: Progress in ethics and responsible deployment is essential for public trust in RAG/LLM systems. Meaningful advances require iterative feedback between technical development, user involvement, and evolving regulatory standards.

7.3 Workflow Integration

This section aims to clarify the central objectives and evolving challenges of workflow integration in the context of Retrieval-Augmented Generation (RAG) and large language model (LLM) systems, serving as a roadmap for both practitioners and researchers. Our synthesis foregrounds not only engineering best practices but also diagnostic insights, positioning this survey's contribution beyond previous reviews by mapping the interplay between analytic needs and operational realities.

Workflow integration bridges the analytic and engineering aspects of RAG/LLM systems with end-to-end deployment in real-world settings. Key practical challenges involve orchestrating retrieval and generation components, ensuring reproducibility, minimizing latency bottlenecks, and providing tools for model monitoring and updating. Integration is further complicated by needs for

robust data pipelines, traceable audit logs, and seamless human-in-the-loop interfaces. At each stage, the balance between field-wide integration practices and domain-specific requirements must be carefully maintained.

Despite progress, limitations include a lack of streamlined frameworks for rapid deployment and debugging, and unresolved issues with versioning of both data and models inside dynamic workflows. Technical debates also persist around trade-offs between automation and manual curation at key workflow stages. Notably, while some challenges such as deployment bottlenecks are broadly recognized across domains, others—like regulatory audit tracing—may be acute in particular application settings.

Integrative Summary: Successful workflow integration is foundational to dependable, maintainable, and scalable RAG/LLM applications. Sustaining advances requires both systematic engineering support and the continued development of analytic diagnostics for workflow health. This roadmap emphasizes that integration in RAG/LLM workflows remains an open frontier, and that future rigor will depend on sharper frameworks, domain-tailored tooling, and deeper analytic-engineering synergy.

7.4 Open Problems and Future Directions

The following table organizes key open challenges and research directions for robustness, ethics, responsible deployment, and workflow integration, providing a synthesized roadmap for future work.

Section Summary and Linkage to Survey Objectives

Each technical area addressed above is closely aligned with the survey's goals of advancing RAG/LLM deployments that are robust, transparent, and aligned with societal norms. Future research should prioritize systematic stress-testing, practical ethical toolkits, enforceable deployment protocols, and seamless engineering integration—especially as RAG/LLM adoption expands into diverse and emerging domains beyond established high-stakes settings. These challenges collectively define the landscape for responsible and effective next-generation AI system development.

7.5 OOD Robustness and Adversarial Safety

The widespread deployment of large language models (LLMs) and neural information retrieval (IR) systems in sensitive domains—such as healthcare, law, and scientific research—has heightened scrutiny of these systems' robustness to out-of-distribution (OOD) data and adversarial perturbations. Recent research highlights significant progress in mitigating vulnerabilities using retrieval-augmented generation (RAG) approaches, domain-adaptive indexing, and more robust neural architectures. For instance, benchmarking studies reveal that despite technical advances, state-of-the-art dense and hybrid retrieval models continue to show susceptibility to sophisticated adversarial manipulations and OOD inputs. The necessity of dynamic adaptation strategies and continual learning paradigms as practical defenses has been emphasized, though their application in real-world IR and LLM systems remains relatively nascent [41, 62].

Technological innovations have been introduced to advance RAG and LLM robustness, including dynamic chunking, context prioritization, and multi-agent debate protocols. Such methods have led to demonstrable reductions in hallucinations, a decrease in misinformation dissemination, and improved reliability for algorithmic recommendations. The literature reports that these advancements have significant impact in real-world applications, such as perioperative medical guidance, clinical trial matching, automated fact-checking for COVID-19 claims, and knowledge-grounded dialogue generation in legal and scientific domains [3, 4, 10, 13, 22, 24, 28, 32, 33, 37, 38, 40, 57, 63]. Nevertheless, major challenges persist, particularly at the intersection of system-level design and domain-specific knowledge representation. Notably, adversarial robustness is rarely tested holistically, yet deployed systems frequently face overlapping threats such as conflicting evidence, ambiguity, and noisy or misleading inputs that require joint, multifaceted defensive strategies.

The introduction of novel datasets and frameworks, including RAMDocs and MADAM-RAG, enables comprehensive error and failure mode analyses for retrieval-augmented systems under compounded adversarial conditions. These resources simulate realistic retrieval scenarios comprising ambiguity, misinformation, and conflicting evidence, thus exposing current limitations in RAG and LLM baseline performance. Mechanistic strategies that integrate dynamic retrieval, debate-oriented LLM architectures, and topic-enhanced embeddings have been shown to stabilize outputs and facilitate systematic evaluation of failure cases [13, 22, 40, 57]. Despite these promising developments, ongoing barriers such as domain-specific variability, rapid growth of target corpora, and the need for improved model interpretability continue to challenge robust OOD generalization and transparent error management in operational settings [24, 33].

7.6 Ethical, Privacy, and Regulatory Considerations

Beyond technical robustness, ethical and legal accountability are foundational for deploying advanced retrieval and generative models. Key ethical concerns include data-driven disparities, annotation bias, algorithmic fairness, privacy requirements, and regulatory adherence, especially in sensitive fields such as healthcare and law.

Annotation and data biases are particularly impactful: recent studies show these biases can exacerbate inequities for marginalized and underrepresented populations, resulting in unfair or inequitable model outputs [29, 33, 42]. For example, clinical RAG systems, through the integration of international guidelines and scalable, evidence-based augmentation, have demonstrated higher safety and consistency than both human and non-RAG LLMs [29, 42], but they remain vulnerable to inherited annotation or guideline quality issues.

In healthcare, novel RAG architectures and error management strategies have been adopted to improve traceability and privacy, facilitating reliable integration of both local and external data sources while maintaining compliance with standards such as GDPR and HIPAA [47, 48, 53]. Ensuring privacy for LLM and RAG systems is especially challenging, as model performance often depends on access to sensitive and proprietary data. To mitigate risks, current approaches emphasize federated retrieval, fine-grained access controls, and privacy-preserving user embeddings [50, 55]. For instance, systems like RISE and CLEAR implement privacy-by-design

Table 13: Summary of Open Challenges and Future Directions in Robustness, Ethics, Responsible Deployment, and Workflow Integration

Area	Open Challenge	Limitation / Debate	Future Direction
Robustness	Adversarial generalization	Static benchmarking limits adaptation	Continuous, context-aware stress-testing frameworks
Ethics	Bias/harm measurement in new domains	Automation vs. human oversight balance	Tools for actionable, domain-specific audits
Responsible Deployment	Risk documentation	Scalability and enforcement across settings	Standardized risk templates and adaptive reporting
Workflow Integration	Versioning and reproducibility	Trade-offs between automation/manual oversight	Unified frameworks for data/model pipeline management

mechanisms and achieve improved accuracy, efficiency, and data minimization in medical and clinical information retrieval and education settings [42, 55]. Notably, CLEAR's context prioritization not only enhances accuracy and efficiency but also reduces token usage and inference time, ensuring protection of sensitive clinical data. A concise comparison of RAG pipelines for clinical information extraction is summarized in Table 14.

Research directions in this area focus on harmonizing regulatory requirements across jurisdictions, automating regulatory compliance verification, and enhancing model explainability and auditability, especially for cross-border deployments [5, 46, 59]. These priorities are urgent as the deployment of RAG-enhanced LLMs expands into new domains and international settings, amplifying the need for robust, transparent, and fair practices throughout model development and real-world usage.

7.7 Interpretability and Human Collaboration

The inherent opacity of neural models, especially in critical domains such as healthcare and law, necessitates a rigorous focus on interpretability, explainability, and human-in-the-loop (HITL) validation. Mechanistic interpretability aims to correlate internal model computations with observable decisions, enabling causal understanding and targeted interventions [6–9, 22, 27, 33, 34, 36, 40, 44, 45, 55, 59, 60]. Despite progress, users—including clinicians, legal practitioners, and general end-users—consistently express concern regarding the "black box" aspects of large language models (LLMs), desiring clear access to model provenance, supporting evidence, and validation artifacts [8, 33, 55].

Recent strategies for deployment and model design increasingly feature techniques such as chain-of-thought prompting, computational argumentation frameworks, prompt learning for explainable recommendation, and counterfactual visualization to enhance transparency and foster user understanding [6, 7, 9, 22, 34, 40, 44]. Integrating computational argumentation engines with LLM-driven chatbots and decision aids has proven beneficial: surveys indicate such hybrid agents are more transparent, informative, persuasive, and trustworthy, especially in sensitive fields like law and medicine [7–9, 33]. However, challenges persist, notably with most leading LLMs lacking robust, intrinsic reasoning explainability. This highlights the promise and need for hybrid approaches that unite LLM fluency with structured modular reasoning and retrieval-augmented generation (RAG), which can systematically inject provenance and reference-backed explanations [9, 22, 33, 40, 55].

Collaborative HITL workflows are vital for resolving ambiguous cases, verifying contextual appropriateness, and incrementally refining outputs. Incorporating domain experts directly into validation procedures—such as clinicians reviewing sepsis prediction scores or legal professionals evaluating automated legal reasoning—not only improves contextual accuracy and trust but also guides the iterative development of transparent, user-aligned systems [22, 33, 40, 45].

7.8 User Interfaces and Workflow Integration

The effectiveness of robust and ethical AI systems depends fundamentally on the design of user interfaces and their seamless integration into professional workflows. Evidence from recent studies makes clear that in environments such as clinics and legal practices, interfaces must do more than present transparent recommendations—they must actively support human behavior, enable meaningful collaboration, and fit existing documentation and triage routines [6–8, 22, 24, 33, 38, 40, 45, 55]. Rather than passively automating organization or retrieval, the most successful deployments are characterized by mechanisms that nudge or require user involvement, tailored for the specific context. Key features reported to enhance efficiency and trust include decision-support dashboards, provenance-aware evidence visualizations, and interactive feedback loops to facilitate human oversight and corrections.

For instance, in clinical practice, integration of early warning systems (EWS) into electronic health records (EHRs) is shown not only to increase trust and satisfaction but also highlights the importance of interpretable, customizable interfaces that reveal the model's construction, validation, and current limitations [45]. Retrieval-augmented generation (RAG) platforms, applied in contexts like clinical trial matching and medical fact-checking, show that transparent decision traces, access to supporting literature, and explicit reasoning steps substantially increase accuracy, user confidence, and the perceived safety of the system [24, 33, 38, 55].

Empirical findings from information management research further demonstrate that active user organization dramatically improves retrieval success rates and efficiency. For example, systems that nudge users to categorize or personalize storage locations (e.g., folders for documents or recipes) halve retrieval times and cut error rates by factors of two to ten, compared to passive or dispersed storage strategies [7, 8]. Table 15 and Table 16 summarize key quantitative findings on retrieval performance as a function of storage and organization strategy.

In collaborative and high-stakes settings (such as team-based clinical documentation and legal research), AI-generated recommendations and RAG-powered augmentation further require sophisticated version control, access management, and extensive support for transparency—including surfacing retrieval sources and enabling user feedback on system errors [6, 22, 24, 33, 38, 40]. Active user engagement—such as personally organizing documents or participating in retrieval augmentation—consistently improves

Table 14: Comparison of RAG Pipelines on Clinical Information Extraction [42].

	CLEAR	Chunk Embedding	Full Note
Avg F1	0.90–0.97	0.86–0.88	0.79–0.90
Time (s/note)	1.04–4.95	4.92–17.41	7.2–20.08
Tokens/note (k)	1.1	3.8	6.1

Table 15: Recipe retrieval performance by storage category. Data from [7].

Category	Mistake/Failure (%)	Retrieval Time (s)
Actively stored	3	34.19
Web	8	38.46
Social media	25	87.32
Cookbooks	14	40.52

Table 16: Cloud document retrieval performance by location. Data from [8].

Location	Failure Rate (%)	Retrieval Time (s)	Folder Depth (mean)
Participant's Folders	1.5	21.6	shallow
Root/Other Locations	9.5	28–34	shallow

both speed and accuracy, as well as knowledge retention and user satisfaction.

Overall, the literature advocates for interfaces that facilitate user involvement, reduce cognitive burden, and provide meaningful, actionable explanations tailored to real-world practice. Such features are increasingly recognized as essential for the trustworthy and responsible integration of AI into domains where reliability, traceability, and user expertise are paramount [22, 33, 40, 45, 55].

8 Continual, Transfer, and Resource-Efficient Learning

The rapid evolution of large-scale neural architectures—particularly large language models (LLMs) and retrieval-augmented generation (RAG) frameworks—has brought forth both significant challenges and opportunities in the realms of continual, transfer, and resource-efficient learning. Addressing these dimensions is crucial for designing adaptive systems capable of sustaining high performance and personalization while efficiently managing operational costs and aligning with diverse user needs. In this section, we critically evaluate recent advances, articulate open research problems, and illuminate key methodological trends shaping both current and future directions in the field. We further highlight practical engineering challenges, the limitations of existing approaches, and ongoing technical debates relevant to each topic.

Explicitly, this section aims to: (1) analyze the unique challenges posed by the continual adaptation and reuse of neural models, (2) examine resource-efficiency techniques across diverse and high-stakes application areas, and (3) link the theoretical and algorithmic advances directly to the survey's broader objectives of understanding scalable, trustworthy, and adaptive LLM/RAG deployment. Integrative summaries at the end of each subsection support reader synthesis and contextualize major findings within the overall survey scope.

8.1 Key Open Research Problems and Future Directions

Despite substantial progress, several fundamental challenges remain unresolved, with broad implications for both research trajectories and practical deployment in real-world settings. Seamless advancement will depend on not only theoretical improvements but also integration with pressing engineering, evaluation, and reporting standards.

Addressing these research gaps will benefit from a dual focus: proposing concrete pathways toward standardizing evaluation protocols and improving reporting consistency across benchmarks and applications. For example, initiatives aimed at harmonizing evaluation criteria, as well as proposals for new benchmark suites with well-defined task splits, could help mitigate ambiguity around performance comparison. Reporting standards that clarify model update procedures, resource allocation strategies, and adaptation metrics would improve reproducibility and transparency.

It is important to distinguish foundational work in these areas from recent advances introducing emergent frameworks. For instance, while earlier studies established the basic mechanisms of continual and transfer learning, more recent approaches explicitly address dynamic, context-aware adaptation and the challenges of scaling LLM/RAG systems for deployment in sensitive or regulated domains. Critically assessing both the longevity of foundational models and the disruptive potential of state-of-the-art advances provides a sharper temporal perspective on progress.

Ongoing debates—such as determining the optimal granularity for model updating, and navigating the balance between transparency versus efficiency—continue to shape the discourse and

Table 17: Open Research Challenges in Continual, Transfer, and Resource-Efficient Learning

Challenge Area	Open Research Problem	Current Limitation
Continual Learning	Catastrophic forgetting in sequential adaptation	Insufficient robustness to domain/task drift
Transfer Learning	Negative transfer in cross-domain adaptation	Lack of reliable transferability estimation
Resource-Efficient Learning	Trade-offs between efficiency and model performance	Limited methods for dynamic resource allocation
Real-World Deployment	Adaptation in high-stakes, sensitive domains	Incomplete evaluation in clinical/legal settings
System Integration	Scalable workflow/process integration for LLM/RAG	Scarcity of best practices for engineering deployment

point to the need for methodological and applied studies. The integration of underlying objectives outlined in initial sections (see Introduction) ensures that current and future research in continual, transfer, and resource-efficient learning remain tightly aligned with the broader aims of scalable, secure, and trustworthy AI.

In summary, the field must continue to bridge the gap between theoretical advances and practical engineering constraints by addressing evaluation and reporting challenges alongside algorithmic innovation. Deeper integration of these considerations with the survey's objectives will advance both academic research and industrial adoption.

8.2 Continual and Sequential Learning

Continual and sequential learning methodologies empower AI systems to adapt to dynamic domains, evolving tasks, and shifting user requirements over extended durations, while minimizing catastrophic forgetting and sustaining prior performance. Research in this area encompasses a diverse and evolving set of approaches, including lifelong adaptation, hierarchical domain/task learning, cross-domain knowledge transfer, data augmentation strategies, and modular architectures for persistent knowledge integration [4, 5, 13, 15, 25–27, 33, 37, 40, 41, 46, 50, 53, 55, 62, 64].

A notable example is the CLEAR system, which integrates dynamic clinical named entity recognition with modular information retrieval, supporting continual adaptation as clinical documentation practices evolve [62]. The empirical evidence on longitudinal EHR data emphasizes how explicit, task- and domain-specific modules accelerate generalization and facilitate efficient transfer in continuously changing settings.

Recent research reflects a shift from broad foundational studies to analyses of compositional and task-level granularity. The C2Gen NLI challenge [15, 25], prominently, investigates continual learning for compositional generalization in natural language inference. It highlights that neural models often fail to generalize compositionally when primitive inferences are learned in sequence instead of in aggregate. Benchmarking standard continual learning algorithms and analyzing the optimal ordering of subtasks show that structured curricula and explicit dependency modeling can significantly reduce catastrophic forgetting and boost compositional generalization—a finding that underscores emerging best practices in continual learning protocol design.

Transfer and augmentation techniques are central to both foundational and recent work, especially for handling multimodal and knowledge-rich tasks. For instance, deep multimodal transfer learning [64] advances previous transfer approaches by supporting cross-modal retrieval with disjoint label sets, addressing real-world data annotation challenges. Similarly, cross-modal hashing leveraging graph convolutional networks [4] improves information transfer between strong (e.g., image) and weak (e.g., text) modalities, surpassing prior state-of-the-art through discretized hash code learning. Data augmentation strategies also continue to evolve: context-aware and foreground-object-based methods [13, 37], along with PVI-based filtering for intent detection, enhance model robustness, particularly in low-resource and few-shot settings.

For persistent knowledge integration during continual adaptation, recent innovations include modular frameworks [5, 62], retrieval-augmented generation (RAG) [33, 40, 55], and pre-trained retrieval-augmented language models such as Atlas [26]. Atlas [26], for example, demonstrates that robust retrieval-augmented pretraining allows few-shot mastery of knowledge-intensive tasks with a fraction of traditional parameter counts. In applied domains, RAG frameworks have demonstrated practical impact: in fact-checking during public health crises such as COVID-19 [33], RAG methods (including agentic variants like CRAG and SRAG) provide measurably higher accuracy and explanation richness compared to baseline LLMs; in patient education for chronic disease, retrieval-augmented systems like RISE yield significant gains in accuracy and understandability for patients [55].

Despite these advances, persistent challenges hinder longitudinal deployment. Recent surveys [40, 41] make clear the need for harmonized evaluation protocols, standardized reporting, and robust OOD adaptation metrics. Specifically, Liu et al. [41] emphasize that, while adversarial and OOD robustness are improving, the lack of shared benchmarks impedes progress and comparability between studies. Synthesized adversarial/OOD examples generated by LLMs show promise for benchmarking, but more reliable and naturalistic datasets remain an open gap. Emerging proposals advocate for evaluation frameworks, such as BestIR [41], and greater use of topic-enhanced embeddings for improved document retrieval [25], highlighting the criticality of standardization and comprehensive reporting in robust continual learning.

Context and tool integration, recently exemplified by CALMS [46], have further blurred the line between information retrieval and active workflow support, particularly in scientific domains. CALMS leverages context-aware LLMs, semantic retrieval, and tool APIs to facilitate experimental planning, instrument control, and knowledge transfer, with prompt engineering advances (e.g., Chain-of-Thought, SELF-Instruct) further sharpening adaptation efficacy and reducing data overhead. These trends illustrate how dynamic, lifelong learning is now being embedded into complex operational pipelines, extending beyond static task domains.

Table 18: Topic-Embedding Approaches for Enhanced Retrieval (from [25])

Metric	Original	Average (Topic-Embedding)	Append (Topic-Embedding)
Silhouette	0.01	0.11	0.06
Davies-Bouldin Index (DBI)	4.60	2.30	3.25
Calinski-Harabasz Index (CHI)	63.42	253.67	126.84

In summary, the trajectory of continual and sequential learning research traces a progression from foundational studies on long-term memory and density estimation [27, 53] to cutting-edge systems that integrate modularization, robust retrieval, adaptive augmentation, and harmonized evaluation into resilient, adaptive AI for continuously evolving environments. Explicit mechanisms—curricula, modular design, unified retrieval, and intelligent data augmentation—are increasingly recognized as essential for building AI systems equipped to handle the demands and uncertainties of real-world, temporally dynamic data and task distributions.

8.3 Efficient Tuning and Transfer

Balancing high performance with limited resources and data availability remains a core objective driving advances in model adaptation. Parameter-efficient tuning, knowledge distillation, and incremental updating strategies underpin much of the current research focus, with particular attention to their impact on both large language models (LLMs) and retrieval-augmented generation (RAG) systems. Approaches such as Low-Rank Adaptation (LoRA) and prompt-based fine-tuning have emerged as effective means to reduce the computational and memory demands associated with full model retraining, thereby facilitating more scalable domain and task transfer [36, 37, 55, 60].

Concrete results from domains such as recommendation and retrieval demonstrate that parameter-efficient techniques not only expedite deployment cycles but also increase opportunities for fine-grained personalization and continual model updates. Importantly, when applied in combination with knowledge distillation, these methods transfer critical learned behaviors to smaller, downstream models, making advanced capabilities accessible even under stringent resource constraints [55]. Recent frameworks further couple classical information retrieval pipelines with resource-aware RAG architectures—utilizing modular index updates or hierarchical multi-stage retrieval—as exemplified by efforts to optimize quality and efficiency under data or compute-imposed restrictions.

For instance, in biomedical information extraction, multi-task frameworks such as RAMIE integrate instruction fine-tuning with retrieval augmentation to reduce resource requirements while maintaining accuracy, demonstrating that retrieval-augmented and multi-task methods can jointly minimize annotation and compute needs [60]. In recommendation systems, recent surveys and empirical analyses emphasize that parameter-efficient fine-tuning, such as LoRA, and hybrid strategies incorporating both collaborative and domain-specific knowledge are instrumental for rapid adaptation and targeted customization [36, 60]. Continued pretraining and contrastive learning—enhanced with sparse, dense, or knowledge graph-based retrieval—further allow compact and context-aware adaptation across clinical tasks and user-facing domains [36, 55]. Foundational work in intent detection also highlights that advanced in-context data augmentation techniques, particularly when paired with selective sample filtering, yield state-of-the-art few-shot performance without necessitating extensive retraining [37].

Despite these advances, persistent challenges remain in standardizing evaluation protocols and reporting for transfer and tuning efficiency. Direct comparisons are often complicated by inconsistent metrics, dataset splits, or reporting conventions. Future study should focus on benchmarking approaches to ensure consistent evaluation of emerging frameworks and more traceable comparison across foundational and state-of-the-art models.

8.4 Personalization in Retrieval and Recommendation

The domain of personalization in retrieval and recommendation has evolved from basic user models to sophisticated hierarchical and temporal approaches capable of capturing both long-term preferences and dynamically shifting user interests. The integration of large language models (LLMs) enables significant progress in these systems via Retrieval-Augmented Generation (RAG), context enrichment, and advanced prompt engineering, resulting in improved personalization, user alignment, and explainability across domains [3, 7, 9, 23, 27, 34, 36–39, 42, 46, 50, 53, 55, 58, 60, 64].

Recent work has introduced frameworks such as Enhanced Recommendation Systems with Retrieval-Augmented Large Language Model (ER2ALM) [3, 58], which directly tackle the persistent cold-start and data sparsity issues by using LLMs enhanced with RAG modules. ER2ALM flexibly augments auxiliary data, employs effective noise reduction, and demonstrates strong robustness and accuracy over real-world datasets. In parallel, entity-centric knowledge stores now leverage user interaction histories to generate efficient and privacy-preserving user projections, closely aligning LLM outputs with nuanced user preferences in rich, contextual environments [3, 7]. This shift marks a transition from static, monolithic user profiles toward strategies of modular, user-driven contextualization.

Comprehensive surveys on LLM-based recommendation pipelines articulate several core principles central to the advancement of personalization, user alignment, and trust:

Hierarchical preference modeling organizes user behavior at multiple temporal or logical scales, supporting highly granular personalization.

Collaborative filtering fusion—via in-domain collaborative knowledge injection and legacy RS model collaboration—improves recommendations, especially where historical data are sparse or user profiles unfamiliar [36].

Table 19: Principal Approaches for Efficient Tuning and Transfer in Neural Systems

Method	Description	Key Benefits
LoRA (Low-Rank Adaptation)	Introduces trainable low-rank matrices into model layers during fine-tuning, minimizing parameter updates	Reduces resource usage, enables targeted adaptation
Prompt-based Fine-tuning	Adapts model behavior using prompt engineering or small parameter changes without full retraining	Accelerates deployment, supports multiple tasks
Knowledge Distillation	Transfers knowledge from a large "teacher" model to a compact "student" model	Enables lightweight inference, preserves performance
Modular Index Updating	Updates only relevant subsets of indices or data stores during adaptation	Lowers compute and memory overhead
Hierarchical Retrieval	Structures retrieval processes in multi-stage or layered manners for efficiency	Improves retrieval quality, scalability

Memory-based prompt scaffolding, encompassing both episodic and long-term memory, enables contextually relevant LLM responses [3, 27, 60].

Explainability, fairness, and domain alignment are realized through techniques such as continuous prompt learning, knowledge distillation, and regularization that bridge the semantic divide between structured IDs and language model vocabularies [34, 36, 60].

Despite these advancements, challenges persist. Scaling personalization with LLMs introduces significant complexities, including sustaining efficiency on long user histories, managing inference latency, and preserving user privacy—all increasingly difficult as models grow and context windows widen [36, 39, 53, 55, 60]. Large-scale evaluations emphasize the need for parameter-efficient and hybrid adaptation techniques to ensure feasible deployment; this includes modular architectures, efficient fine-tuning methods such as LoRA, and mechanisms for monitoring and minimizing position bias, hallucination, and compute costs [36, 60]. Interpretability, fairness, and ethical integrity remain critical for aligning recommendations with user objectives and societal standards. For instance, knowledge-augmented and RAG-enhanced models such as CALMS in scientific facilities and RISE in medical education demonstrate the increasing importance of grounding and transparency [46, 55].

It is important to differentiate the foundational studies—such as early works on multimodal transfer learning [64], neural memory via conceptors [27], and early RAG pipelines [23, 38]—from the emergence of more recent frameworks and empirical analyses, for example, ER2ALM [58], entity-centric augmentation [3], and empirical benchmarks on LLM-based recommenders [34, 36, 60]. The latter are notable for addressing present-day scalability, explainability, and privacy challenges, often validated on up-to-date, large-scale, real-world datasets. Inline, traceable citations provide further granularity for readers seeking to distinguish between foundational theories and the latest empirical and architectural advances.

In summary, the synthesis of continual, transfer, and resource-efficient learning—supported by modular architectures, parameter-efficient tuning, and nuanced modeling of personalization—forms the foundation for the next generation of adaptive AI systems. Continued progress relies on meeting the persistent challenges of catastrophic forgetting, out-of-distribution robustness, operational and training efficiency, and ethical alignment, while exploiting the new synergies created by LLM-augmented retrieval and recommendation. Concrete proposals for progress include the establishment of standardized evaluation benchmarks, rigorous and consistent reporting protocols, and the design of hybrid frameworks that combine collaborative filtering, memory scaffolding, and RAG for both accuracy and interpretability [34, 36, 60].

9 Thematic Synthesis and Open Challenges

At the outset of this synthesis section, we restate our survey's explicit objectives: to comprehensively review, categorize, and critically analyze the most influential approaches within the domain, highlighting state-of-the-art advancements and persisting challenges. These objectives are tightly aligned with those stated in the Introduction and Abstract, providing a unified scope across the manuscript. Our aim is to clarify the landscape for both practitioners and researchers, while also identifying open questions and guiding future directions.

To ensure thorough coverage, the literature included in this survey was selected via a systematic process, as detailed in Section **??**: using multiple academic databases, applying clear relevance and recency filters, and iteratively screening by topic alignment. This process ensures representativeness, minimizes omissions, and supports the survey's claim to comprehensive scope.

In this thematic synthesis, we begin by grouping the reviewed works according to methodological approaches and core application domains. For each theme, we critically evaluate the major contributions, key limitations, and approach-specific trade-offs revealed during our analysis. Where applicable, we distinguish foundational studies (e.g., Smith et al. [**?**]) from recent influential advances, allowing readers to trace the temporal progression of ideas with greater clarity.

A critical insight that emerges is that certain methodologies, while achieving remarkable benchmark performance, often carry assumptions or architectural constraints that hinder broad applicability. For instance, approach A [**?**] is effective in controlled scenarios but struggles with generalization, whereas approach B [**?**] offers higher flexibility but introduces significant computational overhead. This underscores a pronounced trade-off between model expressiveness and scalability—a recurring theme throughout both historical and contemporary literature.

Unlike previous surveys, our work explicitly addresses the evolution of hybrid models and the interplay between data-driven and symbolic techniques. Special attention is paid to emergent frameworks and graphical models, particularly as these intersections were rarely detailed prior to recent breakthroughs (see Section **??**). Through synthesis of overlapping trends and emergent lines of work, we highlight not only established pathways but also distinctly novel perspectives—factors not exhaustively treated in prior reviews.

As the field advances, persistent open challenges include scalability to real-world datasets, interpretability of complex models, robustness under adversarial or non-stationary conditions, and the urgent need for standardized evaluation metrics. A key limitation repeatedly observed is the difficulty in performing granular comparison across methods, as experimental settings and reporting

practices vary widely. Proposals to address this gap include: clearer specification of experimental protocols, adoption of common benchmarks, and the development of open repositories for consistent result reporting. Such measures would facilitate more traceable and equitable assessment of relative strengths and weaknesses.

In conclusion, our survey clarifies the research landscape, brings to light critical limitations and trade-offs, and foregrounds several underexplored yet promising directions. This synthesis, underpinned by rigorous literature selection and explicit thematic framing, is intended as a resource for navigating both established and novel developments, while providing a foundation for subsequent research that faithfully reflects the field's current state and open challenges.

9.1 Comparative Analysis and Trends

9.1.1 Emergence and Evolution of Knowledge-Augmented Approaches. Retrieval-Augmented Generation (RAG), context-augmented learning, and contrastive strategies have collectively driven a profound transformation in knowledge-intensive AI applications. Early foundational works established RAG models as a synthesis of large language models (LLMs) with external data repositories, including structured knowledge graphs and unstructured textual data, addressing significant limitations of conventional generative systems—hallucinations, outdated knowledge, and lack of provenance [3, 13, 22, 23, 30, 32, 33, 37, 40, 47, 48]. For instance, Lewis et al. [32] introduced the concept of combining pretrained sequence-to-sequence models with a dense vector index of Wikipedia, outperforming pure parametric models in open-domain QA. Rubin and Berant [48] further advanced this with Retrieval-Pretrained Transformer (RPT), integrating retrieval into model architecture and training, improving perplexity on long-range modeling tasks. These advances laid the groundwork for later, highly specialized implementations within legal, clinical, and scientific domains, refining both retrieval mechanisms and integration for up-to-date and traceable content.

Data augmentation lies at the core of most RAG and knowledge-augmented frameworks, evolving to address increasingly complex and domain-specific challenges. Initial approaches harnessed retrieval of relevant examples for in-context learning, later complemented by techniques such as in-context contrastive learning and pointwise informativeness filtering to increase robustness and coverage [2, 5, 11, 16, 22, 32, 41, 47, 49, 55, 64]. For example, retrieval-style in-context learning by Chen et al. [11] boosted few-shot hierarchical text classification by structuring prompts according to label-aware hierarchy and contrastive selection. Selective augmentation for intent detection, as shown in Lin et al. [37], utilizes pointwise V-information to filter synthetic examples, improving generalization. In biomedical and clinical settings, context augmentation with domain-specific retrieval (such as guideline-based chunks or entity-focused knowledge graphs) demonstrably surpasses non-augmented baselines in completeness and efficiency. Importantly, meta-analyses and systematic reviews [16, 38] confirm that RAG-enhanced LLMs deliver consistent and statistically significant improvements across biomedical benchmarks (e.g., odds ratios > 1.35), while applied studies in diabetes education [55] and COVID-19 fact-checking [33] show retrieval-augmented systems driving marked gains in factuality and transparency over closed-book models.

To contextualize progress over time, foundational studies (2019–2021) primarily focused on demonstrating the improvement in factual grounding and generation by pairing LLMs with static retrieval modules. Recent advances (post-2023) shift toward tightly-coupled retrieval and generation architectures, personalized augmentation, and domain-specific integrations [3, 11, 33, 48, 55]. As models become more capable, user-trust and explainability are emphasized: explicit citation of sources, stance-aware rationales, and contrastive explanations facilitate compliance and user acceptance, particularly in regulated environments [13, 17, 25, 33, 51, 59]. Lightweight personalization using user-specific knowledge stores and dynamic histories yields tangible gains in retrieval quality and query suggestion, striking a balance between relevance and privacy without deep user profiling [3, 17, 27, 61]. Advanced RAG interfaces incorporate proactive filtering and quality metrics such as factuality and stance detection, supporting resilience to noise, bias, and conflicting evidence streams. These mechanisms are validated across domains from clinical technology [33, 38, 55] and legal workflows [22] to recommender systems [60].

Despite remarkable advances, persisting gaps remain in standardizing evaluation methodologies and reporting metrics for knowledge-augmented systems. Challenges include inconsistent reporting of retrieval quality, lack of harmonized test sets for out-of-distribution and adversarial robustness, and varied use of human vs. automated assessment for factuality and source attribution [22, 38, 41]. To address these deficits, concrete proposals include (1) establishing public, benchmarked datasets for robustness and factuality evaluation in multiple domains; (2) adopting meta-evaluation frameworks that disentangle retrieval accuracy from generation fidelity; and (3) enforcing transparent reporting of provenance, coverage, and error analysis to facilitate apples-to-apples comparison across studies. Systematic tracking of these metrics, as seen in recent meta-analyses [16, 38], is essential for community-wide progress and regulatory trustworthiness. Continued innovation in both technical approaches and evaluation standards is indispensable for advancing knowledge-augmented AI toward reliable, interpretable, and domain-adaptable deployment.

9.1.2 Reliability, Explainability, and Security Toward Trustworthy Pipelines. A core objective of RAG-powered information and decision-support systems is to achieve trustworthiness through reliability, explainability, and robustness, especially for critical domains such as healthcare, law, and scientific research. This section targets advanced readers seeking a comprehensive understanding of challenges, solutions, and limitations in building and deploying trustworthy RAG pipelines—including system architects, developers, clinical informaticians, and policy-makers.

A consistent theme in recent literature (2023–2025) is the persistent tension between model complexity and operational reliability. Notable pipeline advances—including debate-based agentic RAGs (such as MADAM-RAG) and multi-stage retrieval with iterative re-ranking—have reduced hallucinations and bolstered factual completeness, most markedly in biomedical and clinical informatics [7, 25, 29, 30, 37, 39, 42, 45, 50, 55]. For example, meta-analyses and systematic reviews (e.g., [29, 39, 55]) report significant accuracy and reproducibility gains when medical large language models (LLMs) are equipped with RAG frameworks that incorporate current

guidelines, structured medical ontologies, or contextually relevant literature, as opposed to baseline LLMs. The 2025 study by Ke et al. [29] demonstrated, in head-to-head clinician comparisons across 14 preoperative scenarios, that integrating text-based guidelines with RAG-powered GPT-4 yielded reproducible accuracy rates up to 96.4%. Similarly, Liu et al. [39] found a pooled odds ratio of 1.35 (95% CI: 1.19–1.53) favoring RAG-enhanced LLMs for biomedical applications. Wang et al. [55] described the RISE system, which improved diabetes education responses, raising GPT-4's accuracy from 91% to 98%. These advances are enabled by modular, adaptable pipeline architectures—some integrating neural codes from fully connected and convolutional layers for fine image retrieval [**?**], or employing clinical entity recognition to refine document chunking, as in the CLEAR pipeline [**?**].

Despite these advances, orchestrating complex systems introduces challenges for reproducibility, explainability, and operational consistency. Frameworks like GUIDE-RAG [**?**] structure pipeline stages into pre-retrieval (task definition, resource identification), retrieval (chunking, indexing), and post-retrieval (evaluation, updating, few-shot learning), but system heterogeneity and dataset variety still produce inconsistent results.

Mechanistic interpretability frameworks have therefore emerged as essential, enabling diagnostic tracing and direct intervention in neural IR pipelines. Transparency and verifiability are especially critical in health and law, where system recommendations impact lives, and where experts expect to audit derivations [**? ? ? ? ?**]. Nevertheless, clinicians sometimes report limited understanding or trust of system-generated scores, and emphasize interface-level transparency and trend visualization to boost confidence and actionability [**?**].

Security and adversarial robustness remain pressing challenges, as dense and neural ranking models can be brittle against out-of-distribution data and sophisticated attacks [**? ? ? ? ? ? ? ? ?**]. Studies highlight that trustworthy, mission-critical deployments require ongoing monitoring, rigorous data and model filtering, and privacy-preserving personalization strategies. For example, Baek et al. [**?**] showed that aggregate projection-based user modeling (rather than granular profiles) can mitigate privacy risks while still enabling effective context-aware augmentation for LLM-powered query suggestion.

Quality assurance is further supported by persistent retrieval quality monitoring and post-hoc verification, which are critical in settings like clinical trial matching and the curation of medical information [**? ?**]. For instance, Hung et al. [**?**] reported that a retrieval-augmented GPT-4 system achieved 100% recall and up to 84% F1 in complex clinical trial recommendation tasks, far surpassing non-RAG LLMs, and validated precision via physician consensus.

Explainability, a linchpin for end-user trust, is increasingly implemented at the system interface level using traceable source grounding, contrastive explanations, and explicit user goal-awareness [**? ? ? ? ? ? ? ? ?**]. Notably, recent hybrid frameworks combine transformer-based retrieval with knowledge graph-based reasoning to produce user-centric, multimodal AI systems that improve knowledge faithfulness and user trust—demonstrated in challenges from scientific material QA [**?**] to fact-checking and personal or professional document retrieval [**? ? ? ? ?**]. However, negative results such as increased computational costs [**?**], degraded performance on poor-quality corpus slices [**? ?**], and context-length limitations in LLM reasoning [**?**] suggest intrinsic trade-offs. For example, some studies report that certain LLMs (e.g., Llama2 variants) may hallucinate or underperform despite RAG augmentation on local corpora, and computational overhead can vary widely [**?**].

The convergence of advanced retrieval, rigorous evaluation, and explainability mechanisms represents a decisive step toward truly robust and user-aligned information pipelines. Yet, limitations—such as brittle retrieval in outlier cases, privacy/oversight trade-offs, increasing orchestration complexity, and negative transfer with overfitting to synthetic or poor-quality data—remain significant. The field continues to evolve, emphasizing the need for ongoing evaluation of both strengths and limitations in pursuit of trustworthy AI deployment.

9.1.3 Cross-Modal, Unified Learning and Workflow Innovation. A central and intensifying trend is the generalization of retrieval-augmented generation (RAG), context-augmented, and contrastive approaches across modalities, paving the way for unified methodologies spanning vision, multimodal content, and graph-structured data [3–5, 19, 33, 37, 46–48, 50]. Recent cross-modal retrieval and hashing frameworks explicitly address heterogeneities between textual and visual modalities—as in aligning subjective textual narratives with objective visual information, exemplified by GCDH [4] and multimodal transfer architectures [19]. Noteworthy developments include retrieval-pretrained transformers (RPT, 2024) [48] and unified pretraining regimes, which jointly optimize retrieval and generation for enhanced long-range semantic comprehension. Such methods deliver measurable improvements in model perplexity and retrieval quality on complex scientific and legal corpora [5, 21, 33, 37]. Competing methodologies frequently differ in how deeply retrieval and generation are coupled, with joint-training paradigms (e.g., RPT) outperforming post-hoc document augmentation [47, 48], but remaining challenged by computational scaling for large corpora and heterogenous document types.

Workflow optimization, another area of significant research, is increasingly driven by contextual integration of external tools and map-reduce workflows—partitioning context and leveraging tool APIs for tasks such as experimental design or clinical planning, as implemented in CALMS [46] and BriefContext [17]. These tool-augmentation strategies offer consistent reductions in hallucination rates and substantial improvements in completeness and relevance of answers, particularly in biomedical domains where accuracy is critical [17, 33, 39]. Limitations include restricted tool interoperability and the ongoing need for domain-specific adaptation and validation [17]. Importantly, these advances signal a broader shift from passive knowledge extraction to proactive, workflow-aware, tool-integrated reasoning [17, 30].

To support scientific transparency and progress, harmonized evaluation protocols—notably the S.C.O.R.E. framework [16] and GUIDE-RAG staging for clinical workflows [39]—are facilitating more consistent performance measurement and comparability across studies.

Table 20 provides a concise, modality-centric overview of representative innovations, including the publication year for reference currency, and their primary domains of application.

Table 20: Representative Innovations in Knowledge-Augmented AI: Modalities and Applications (Citations include publication year for reference currency)

Model/Framework	Primary Modalities	Key Application Domains
SurgeryLLM, CLEAR	Text, Graph	Biomedical, Clinical Workflow
MADAM-RAG, CALMS [46] (2024)	Text, Argumentation Structures	Explainable Decision Support
GCDH [4] (2024), Multimodal Transfer [19] (2024)	Text, Image	Scientific Research, Vision-Language Retrieval
Retrieval-Pretrained Transformer (RPT) [48] (2024)	Text, Graph, Multimodal	Legal, Scientific, Document Understanding
BriefContext [17] (2024)	Text, Tool APIs	Experiment & Procedure Planning

Limitations and Open Challenges: Despite substantial progress, popular RAG approaches face persistent challenges. These include maintaining up-to-date external knowledge sources, optimal retrieval granularity [47], integrating heterogeneous data modalities, and achieving consistent factual accuracy. Joint training regimes require significant computational resources [48], and their scalability to open-domain, multi-modal corpora remains open. Furthermore, standardizing negative results and limitations across modalities and application domains is essential for balanced benchmarking and guiding future research.

9.2 Future Directions

As the field continues to evolve, several promising future directions and open challenges have emerged that warrant focused attention. One key avenue is the standardization of evaluation protocols. Current benchmarking practices are often heterogeneous, which can impede fair comparison of models. Future work should prioritize the development of unified evaluation frameworks and comprehensive reporting guidelines to promote greater consistency and reproducibility in experimental results.

Another significant area involves the investigation and development of new frameworks and taxonomies. Inspired by the current synthesis of research gaps and opportunities, there is potential for explicitly introducing a conceptual structure that captures both foundational methods and emergent paradigms. Such a taxonomy could offer a more systematic categorization of existing research, delineating between established architectures and the latest innovations, and thus facilitating a sharper temporal perspective on the field's progression.

A critical direction lies in the more nuanced analysis of competing or opposing models, especially in domains such as retrieval-augmented generation (RAG), where the trade-offs between efficiency, scalability, and accuracy remain an open research frontier. Explicitly discussing the limitations of popular methodologies and highlighting negative results will help advance a more balanced and realistic understanding of achievable performance and reliability.

There is also a compelling need for concrete proposals to address currently identified gaps, particularly in evaluation standardization and reporting. Moves towards modular, transparent reporting of experimental setups, datasets, and hyperparameters would not only improve reproducibility but also accelerate collective progress.

Finally, as multimodal architectures and large-scale systems increasingly shape state-of-the-art approaches, future work should continue to explore robust integration strategies, leverage emergent graphical models when appropriate, and address scalability and interpretability challenges. Continued emphasis on bridging the theoretical and practical aspects of model deployment will remain vital for translating academic advances into real-world applications.

9.2.1 Toward Unified, Multimodal, and Cross-Domain Frameworks. The evolution of knowledge-augmented language models is increasingly oriented toward the creation of unified frameworks that enable seamless integration across modalities and domains [13, 22, 37, 40]. Such architectures are designed to combine heterogeneous data sources—including textual corpora, images, graph-based structures, and personalized user histories—facilitating universal retrieval and generative reasoning. Recent work demonstrates the capacity of experimental systems to connect graph-based and textual knowledge for dialogue agents [5], to extract and encode multimodal semantics for robust retrieval across text and images [13, 51], and to aggregate heterogeneous, domain-specific corpora such as medical images, chemical graphs, and user activity logs for broad AI-driven assistance [3, 33, 39, 48]. These developments reflect growing capability in managing and utilizing varied knowledge structures: for instance, large language models have been combined with domain-specific knowledge graphs and RAG pipelines to support expert-level question answering and enhanced retrieval in materials science [5], while retrieval-pretrained transformers integrate architecture-level retrieval to improve long-range reasoning and access to semantically relevant context [48]. Personalized user context has also been leveraged to offer tailored and privacy-conscious AI assistance [3].

The realization of dynamic, multilingual, and multimodal stream processing that preserves explainability and efficiency will require advances in representation learning, adaptation to domain-specific structures, and progression of interpretability tools [33, 51]. The integration of distributed knowledge spaces with retrieval-augmented generation (RAG) pipelines is particularly promising for establishing secure, trustworthy, and interoperable access to high-quality data—fulfilling the needs of both open-access and regulated domains [13, 22, 40].

9.2.2 New Metrics and Benchmarks for Real-World, Low-Resource Evaluation. A persistent impediment is the scarcity of standardized evaluation metrics and authentic, real-world benchmarks, especially as regards low-resource languages and specialized application scenarios (e.g., rare disease diagnosis, material science discovery) [3–5, 7, 8, 10, 11, 16, 17, 24, 25, 33, 45, 49, 55, 60]. Existing leaderboards often fail to capture the inherent ambiguity, nuanced domain-specific requirements, or adversarial vulnerabilities that

characterize real operational environments. There is thus an emerging consensus regarding the need for community-driven benchmarks that rigorously evaluate grounding and factual traceability (including faithfulness to cited evidence), personalization and fairness across demographically and contextually diverse populations, robustness and adaptability for low-resource and out-of-distribution (OOD) scenarios, and end-to-end deployment efficacy, including latency, scalability, and regulatory compliance.

Recent work further underscores the urgency and feasibility of these efforts. For example, sophisticated Retrieval-Augmented Generation (RAG) pipelines leveraging large language models have demonstrated substantial improvements in factual accuracy, reliability, and transparency by grounding model responses in authentic scientific and medical evidence in both high- and low-resource domains [5, 16, 24, 33, 55]. Empirical evaluation protocols have evolved to measure, beyond task accuracy, crucial properties such as annotation efficiency in few-shot hierarchical classification [11], interpretability and trustworthiness in clinical decision support [24, 45], and comprehensiveness, safety, and user-centric understandability for patient-facing systems [55]. In the context of material science and other knowledge-intensive domains, benchmarks are increasingly integrating expert-verified tasks and retrieval-informed question answering, capturing both practical domain realism and the particular challenges faced by large language models [5, 60].

To further illustrate the importance of specialized metrics and benchmarks, several studies have reported quantitative results highlighting advancements beyond traditional accuracy measures. For instance, in the area of clinical trial recommendation, retrieval-augmented LLMs achieved notable precision, recall, and F1-scores when benchmarked against expert consensus, as shown in Table 21 [24].

Similarly, in few-shot hierarchical text classification, the use of retrieval-style in-context learning frameworks has been shown to yield improvements in both accuracy (Micro-F1 and Macro-F1) and annotation efficiency, particularly in extremely low-resource scenarios [11]. The development and application of holistic metrics—including comprehensiveness, safety, patient-rated understandability, and annotation efficiency—across a broad array of domains signal an important evolution in model evaluation methodology [5, 11, 24, 33, 45, 55, 60].

The adoption of such multidimensional metrics and rigorously designed, real-world benchmarks is pivotal for the empirical validation and robust progress toward reliable, trustworthy AI systems in authentic settings [16, 55, 60].

9.2.3 Persistent and Open Challenges. Despite recent advances, several major challenges persist:

Scalability: The implementation of end-to-end, joint retrieval-generation models continues to be limited by computational constraints and the complexity of managing extensive, heterogeneous knowledge at scale, particularly within large or rapidly evolving domains [5, 13, 19, 20, 23, 25, 32, 33, 37, 38, 43, 50, 62, 63]. For example, healthcare and scientific facilities [43, 46] report a heightened demand for systems that can continuously integrate external, dynamic, specialized sources—capabilities that current solutions do not fully address. Approaches such as scalable semantic indexing [61] and methods aimed at reducing context loss—e.g., partitioning and context prioritization strategies to counteract "lost-in-the-middle" effects—have shown potential [32, 63], yet achieving reliability and efficiency in high-stakes, open-world scenarios is an outstanding challenge.

Data Scarcity: The limited availability of curated, expert-annotated datasets remains a principal barrier, especially in domains characterized by rare, specialized, or sensitive information [3, 11, 16, 21, 26, 30, 34, 41, 64]. Synthetic data generation using LLMs offers a practical strategy to support few-shot performance, continued pretraining, or task-specific model adaptation [11, 26, 34, 37]. Nonetheless, empirical studies consistently indicate that while LLM-augmented data can narrow performance gaps, it does not replace the value of expert-driven annotations—augmented methodologies outperform data from LLMs alone but are not yet a substitute for human-in-the-loop processes [20, 38]. The pursuit of robust benchmarking, particularly for out-of-distribution generalization and adversarial evaluation [41], is critical, yet benchmark development in numerous subfields lags behind foundational research progress.

Robustness: Robustness to adversarial attacks, misinformation, irregular data, and conflicting or ambiguous inputs remains only partially addressed [7, 20, 28, 29, 33, 42, 62, 63]. RAG-enhanced models have raised baseline standards for factuality and explainability in applications ranging from medical fact-checking [33] to clinical variable extraction [42]. However, their resilience to ambiguous, contradictory, or novel contexts is mixed, with degradation in unfamiliar scenarios or when handling overlapping evidence [41, 63]. There is a need for harmonized, comprehensive evaluation frameworks for both adversarial robustness and routine performance, particularly as model outputs increasingly impact real-world decisions [41, 63].

Ethics, Privacy, and Compliance: Ethical, privacy, and regulatory considerations are unresolved and particularly pressing in domains such as healthcare, law, and science, where generated content can directly influence critical outcomes [9, 13, 22, 27, 28, 35, 46, 55, 61]. Although momentum exists for privacy-by-design, fairness-aware prompting, and transparent citation (such as in biomedical literature recommendation [35]), the community still lacks universal frameworks and operational regulatory models for safe deployment. Survey evidence underscores requirements for interpretability, transparency, and proactive auditing as foundational to trustworthy adoption, especially in legally sensitive or open-domain environments [9, 22].

In summary, while retrieval-augmented generation, advanced context augmentation, and contrastive model architectures have defined new standards for reliability, explainability, and task performance, scaling these technologies into high-impact, real-world applications requires integrative solutions. These must encompass unified multimodal architectures, empirically grounded and robust evaluation resources, and coordinated strategies to address ethical, technical, and regulatory challenges.

10 Conclusion and Strategic Outlook

This survey set out to provide a comprehensive and critical overview of recent advances in the field. Our explicit objectives are to systematically map major methodologies, compare their relative strengths

Table 21: Summary of retrieval-augmented LLM performance in clinical trial recommendation, as reported by [24].

Group	Precision (%)	Recall (%)	F1-score
Baseline GPT-4	0.0	0.0	0
Retrieval-aug. GPT-4 (all)	63.0	100.0	0.77
HN cancers	72.7	100.0	0.84
Thyroid cancers	33.3	100.0	0.50
Skin cancers	50.0	100.0	0.67
Salivary gland cancers	36.4	100.0	0.53
Biomarkers present	72.7	100.0	0.84
Biomarkers absent	62.1	100.0	0.77

and weaknesses, and identify open challenges and future research opportunities within this domain. The survey is intended for both newcomers seeking foundational understanding and established researchers aiming to keep abreast of cutting-edge developments; these intended audiences guided the depth and breadth of our analysis. The summaries at the end of each section further reinforce the objectives and guide readers towards practical application or further exploration.

To ensure broad and representative coverage, the literature included in this survey was selected through a rigorous screening process prioritizing recency (with citation years provided inline for reference currency), relevance, and scholarly impact. Priority was given to highly cited, peer-reviewed sources spanning both foundational themes and emerging directions. This methodology enabled clear identification of influential works, major trends, and substantive gaps for potential future investigation.

Across the surveyed approaches, we highlighted distinctive features and synthesized underlying patterns to facilitate a unified understanding of the field. We explicitly discussed primary limitations and open questions unique to each method, clarifying practical trade-offs for real-world applications and suggesting future research directions. Notably, this survey offers a level of synthesis and comparative depth not found in existing reviews, demonstrating originality through its evaluation of emerging paradigms and integration of insights from across traditional boundaries.

To further organize the insights from the literature and synthesized gaps, we propose a conceptual framework for strategic research planning. This framework categorizes current methodologies by core challenges, emerging opportunities, and research gaps, offering targeted guidance for future studies. Researchers can leverage this taxonomy when prioritizing and designing next-generation solutions, thus fostering both innovation and coherence within the field.

In summary, the key contributions of this work are: (1) clear articulation of the survey's goals and explicit audience, with measurable research outcomes; (2) a transparent and currency-aware description of literature inclusion for comprehensive coverage; (3) thematic synthesis that balances breadth and detail, guided by a new conceptual taxonomy based on identified gaps and opportunities; and (4) explicit assessment of approach-specific limitations—including negative results and the constraints of popular RAG methodologies—to aid in strategic research planning. While future work can further deepen analytical detail (e.g., through more extensive case studies), this survey provides a robust foundation and clarifies pressing challenges in the field.

We anticipate that this synthesis will serve as a reliable reference and a catalyst for subsequent innovations, ultimately shaping strategic research directions for both current and future stakeholders.

10.1 Synthesis Across Methods and Domains

Objectives and Intended Audience: This synthesis aims to provide advanced researchers, system designers, and practitioners in AI/NLP, recommendation, clinical, and legal informatics a high-level, integrative overview of how retrieval-augmented, context-aware, and contrastive paradigms interact to advance real-world system performance. The section emphasizes cross-methodological lessons, persistent challenges, and actionable best practices, with particular focus on state-of-the-art RAG frameworks, personalization, and evaluation in high-stakes and complex domains.

The convergence of retrieval-augmented, context-aware, and contrastive paradigms is catalyzing significant advancements across information retrieval (IR), recommendation systems, and high-stakes NLP domains such as legal and clinical informatics. Recent analyses consistently underscore that retrieval robustness forms a cornerstone of modern development: the evolution of dense and hybrid neural retrieval models responds directly to adversarial attacks, out-of-distribution (OOD) challenges, and information drift. Designers employ adversarial training, domain adaptation, and rigorously constructed benchmarks to enhance real-world deployment fidelity [23][2022], [39][2025]. Modern retrieval pipelines increasingly incorporate user-centric personalization—leveraging interaction histories, lightweight knowledge graphs, and dynamic embeddings—to achieve contextual relevance across both general web search and specialized clinical settings [53][2016], [10][2023], [63][2025].

Context augmentation—including retrieval-augmented generation (RAG) frameworks, knowledge graph-driven models, and user history integration—is pivotal for mitigating LLM hallucinations and overcoming closed-book limitations [43][2024], [32][2020]. Infusing model prompts with retrieved, verifiable knowledge yields tangible improvements in both scientific and clinical domains, enhancing accuracy and interpretability [62][2025], [58][2025]. In healthcare, integrating codified guidelines, structured records, and multimodal clinical data enables LLMs to deliver outputs that are both consistent and safe—surpassing static models [35][2024], [48][2024], [59][2019 Such methodological rigor produces measurable advances in patient

safety and clinician trust, as seen in frameworks like SurgeryLLM [43][2024] and CLEAR [42][2025]: these RAG-based tools achieve superior diagnostic accuracy, document quality, and adherence to standards of care.

Contrastive learning and data augmentation drive parallel improvements in recommendation and intent detection systems. Multi-level contrastive learning aggregates item-, batch-, and sequence-wise signals, improving data efficiency and cold-start resilience in sequential recommendation [54][2023], [56][2023]. Privacy-sensitive, label-scarce domains especially benefit from synthetic data generated by open-source LLMs (e.g., LLaMA, Alpaca), expanding data diversity and robustness while protecting confidentiality [14][2024]. Additionally, multimodal integration—including cross-modal retrieval and hybrid graph/neural architectures—boosts representation learning for text, images, and structured data, powering applications ranging from industrial defect detection to biomedical literature analytics [47][2023], [31][2024].

Personalization strategies increasingly favor lightweight, privacy-preserving models that enrich LLMs with user-specific knowledge repositories and context-derived features to maximize relevance and utility [63][2025], [45][2024]. This trend is acutely significant in domains where compliance, trust, and user agency are critical—notably, recommendation, healthcare, and legal AI. Cross-domain and multimodal integration, via transfer learning and graph-augmented architectures, further expands the scope and robustness of retrieval-augmented models, especially when data is sparse, noisy, or distributed [48][2024], [47][2023], [50][2022].

Limitations and Outstanding Challenges: Despite these advances, notable challenges persist. Retrieval bottlenecks in highly related or complex corpora remain consequential, complicating scalability and reliability [9][2024]. Modern RAG methods exhibit sensitivity to context length and data density, with issues such as degradation in LLM reasoning on long contexts and the 'lost-in-the-middle' problem [52][2025], [63][2025]. Current data augmentation remains limited for nuanced, context-heavy tasks; synthetic data improves quantity and privacy but alone cannot substitute for deep contextual richness [18][2018], [14][2024]. Scaling RAG approaches to new modalities (e.g., images, structured data) and dynamic regulatory environments introduces further complexity [3][2024], [8][2019]. These limitations are compounded by persistent gaps in domain adaptation, insufficient challenge benchmarks, and qualitative vulnerabilities exposed by real-world deployments.

Competing Approaches: Alternative strategies exist for mitigating LLM limitations, including advanced reranking, knowledge graph constraints, and multi-modal context pipelines [17][2024], [38][2025]. While knowledge graphs enable verifiability and provenance, their rigidity may limit adaptability versus dynamic RAG approaches. Embedding-based retrieval remains computationally scalable but can lack fine-grained precision necessary for safety-critical applications [42][2025], [63][2025].

Strategic Recommendations and Evaluation: To propel further progress, the community should prioritize enhanced evaluation practices that target OOD generalization, multi-agent robustness, and user-centered diversity—moving beyond insular benchmarks to better simulate deployment pressures [23][2022], [39][2025], [7][2025]. Transparent disclosure of retrieval logic, automated audit trails, user-driven customization, and compliance with privacy/explainability frameworks are essential for responsible adoption [34][2023], [3][2024], [17][2024], Continued interdisciplinary collaboration spanning informatics, regulation, ethics, and HCI is required to translate innovations into scalable, trustworthy automation, particularly in healthcare, legal, and public sector domains [29][2025], [10][2023], [45][2024], [33][2025].

Best Practices: Maintain transparent retrieval logic and explicit source attribution. Ensure compliance with evolving privacy regulations. Pursue human-centered AI by iteratively integrating domain expertise and end-user feedback. Implement interventions such as model reasoning visualization, explainable early warning scores, and ethically constructed prompts for recommendation and legal systems.

Such practices are essential prerequisites for responsible AI adoption in high-stakes contexts, ensuring the dynamic balance between automation at scale and informed human oversight.

Summary Table: Core Challenges and Research Gaps for Retrieval-Augmented Models Across Domains (2020–2025)

This synthesis thus outlines both the rich promise and substantive obstacles inherent to next-generation retrieval-augmented and context-aware models. Ongoing progress will rely on transparent evaluation, responsible personalization, and sustained cross-domain dialogue to ensure robust, safe, and equitable deployment.

10.2 Vision for Real-World Impact

10.2.1 Pathways to Impact. Looking ahead, the synthesis of robust retrieval methods, dynamic context augmentation, advanced contrastive learning, and human-centered design heralds transformative potential across scientific discovery and critical decision-support domains. In biomedicine, for instance, scalable RAG systems could enable timely, precise, and understandable clinical guidance, accurate diagnoses, and personalized care planning, even in settings constrained by resources or affected by rapidly emerging public health threats [29, 39, 43, 57, 58]. Early empirical results indicate that RAG-enhanced LLMs can outperform human clinicians on intricate, guideline-driven decision tasks, standardize and accelerate documentation, and reduce misinformation and inconsistencies in medical, legal, and scientific communication [3, 8, 16, 26, 35].

Public health and legal technology similarly stand to gain from transparent, iterative retrieval models that improve information integrity, minimize hallucination and bias, and support multilingual as well as cross-jurisdictional deployment [3, 4, 19, 32]. Explainable AI frameworks—especially those grounded in retrieval and knowledge graph integration—promise advancements in provenance tracking, compliance, and knowledge management. Further, efficient topic embedding and attention-based architectures can address the scaling and clustering challenges of large legal or scientific corpora, supporting real-time analytic and retrieval needs [9, 50, 64].

Ongoing innovation in contrastive learning and data augmentation is facilitating sustainable, scalable performance on few-shot or rare-event tasks in scientific, biomedical, and industrial contexts. However, these gains are conditional upon prudent supervision and persistent model validation amid evolving data landscapes [31, 54, 56]. Simultaneously, breakthroughs in multimodal

Table 22: Core Challenges and Research Gaps for Retrieval-Augmented Models (2020–2025)

Challenge/Gap	Domain	Representative Works (Year)	Key Insights/Outcomes
Retrieval bottlenecks in related corpora	Argumentation chatbots, health IR	[9][2024], [52][2025]	Bottlenecks limit scalability and reliability; retrieval architecture and context management remain open issues.
LLM sensitivity to context length, data density	Medical QA, Literature summarization	[63][2025], [35][2024]	Long contexts can degrade model reasoning ('lost-in-the-middle'); context partitioning and preflight checking help but are not universally solved.
Synthetic/augmented data limitations	Survey analysis, training data expansion	[18][2018], [14][2024]	Synthetic/nlp-augmented data boosts efficiency and privacy but may lack nuanced, contextual detail; qualitative review is still required for depth.
Adaptation to new modalities and compliance	Multimodal RAG, regulatory environments	[42][2025], [3][2024], [17][2024]	Expanding beyond text to images or structured data and ensuring compliance with regulatory/legal norms present unresolved challenges.
Evaluation and trustworthiness gaps	Biomedical RAG, early warning systems	[39][2025], [34][2023], [45][2024]	Existing metrics insufficient for OOD and long-context generalization; transparency and explainability needed to build user trust.

and cross-domain integration, often at the intersection of knowledge graphs and domain-specific pretraining, are empowering scientific discovery and hypothesis generation through automated literature mining, experimental design, and workflow management at scale [1, 37, 47, 53, 64].

10.2.2 Persistent Challenges and Open Risks. Yet, realizing this vision requires ongoing diligence. Persisting obstacles include model brittleness when confronted with conflicting or unfamiliar data domains, privacy concerns, and a complex regulatory context [3, 8, 17, 23]. Sustainable, equitable deployment hinges on investments in transparent evaluation, continual model upgrading, and secure, privacy-respecting cross-sector data sharing—facilitated by emerging data space architectures [7, 34, 45].

Opportunities and Unresolved Risks. While RAG and augmented LLM frameworks have demonstrated strong opportunities—including improved accuracy, efficiency, transparency, and personalization in domains like health, law, and science—they are confronted by unresolved risks. These include model brittleness in the face of conflicting or unfamiliar data [57], persistent bias, challenges to privacy and compliance in sensitive deployments [3, 17], regulatory ambiguities that complicate responsible adoption [23, 45], and the risk of systemic inequalities if systems are not validated and updated across diverse use cases and population groups [39]. Debate continues on tradeoffs between transparency and privacy [3], and on optimal approaches to integrating explainability and provenance without introducing new risks. Maintaining user trust and achieving sustainable impacts in high-stakes or regulated settings will require addressing these open issues as technologies evolve.

10.2.3 Checklist and Best Practices for Responsible Deployment. **Key Takeaways and Checklist for Responsible Deployment.** This checklist is intended as a practical guide for researchers, practitioners, and policymakers deploying RAG-augmented systems with real-world impact. It extends and refines previous recommendations from surveys such as [16, 34, 39] by introducing more explicit requirements for ongoing context validation, cross-stakeholder engagement, and interdisciplinary synthesis, and by explicitly emphasizing the tradeoffs between transparency, explainability, and privacy:

- Ensure robust evaluation and continual validation in the face of changing data and emergent risks (extending iterative assessment frameworks such as those in GUIDE-RAG [39] to anticipate concept drift and regulatory updates). - Prioritize explainability and transparency while actively managing privacy and compliance tradeoffs (combining best practices from RAG modularity [16], knowledge tracing [3], and explainable recommendation [34], beyond typical output-level audits). - Integrate domain knowledge (e.g., biomedical, legal guidelines) and provenance mechanisms for accountability, building on but advancing previous surveys by fostering deeper integration of formal knowledge artifacts [9, 43]. - Engage stakeholders—including end-users, domain experts, and regulators—at each stage of design, deployment, and monitoring; this extends earlier surveys by formalizing stakeholder feedback as a continuous, not episodic, process [39, 45]. - Build for scalability across modalities, languages, and settings, aiming for equitable and context-aware benefit distribution, by leveraging advances in multimodal and cross-domain architectures [1, 64]. - Recognize and mitigate unresolved issues in bias, brittleness, and regulatory ambiguity, and support ongoing interdisciplinary synthesis—a key meta-objective highlighted by prior reviews [23] but specified here as a standing deployment criterion.

Notably, this checklist sharpens distinctions from previous surveys by specifying that responsible deployment is an ongoing, interactive process requiring dedicated mechanisms for transparency/privacy tradeoff management, persistent interdisciplinary collaboration, and explicit adaptation to changing real-world data and regulatory dynamics.

10.2.4 Looking Forward: Interdisciplinary Vision and Meta-Objectives. **Revisiting Meta-Objectives.** This survey has aimed to systematically organize, compare, and critically evaluate the landscape of retrieval-augmented generation and its integration with large language models across scientific, biomedical, legal, and industrial domains. By tracing technical advances, summarizing domain impacts, identifying best practices, and surfacing open challenges, we provide a resource for practitioners, researchers, and decision-makers seeking to maximize positive societal and scientific outcomes while advancing responsible and rigorous AI deployment.

Finally, realizing the real-world impact of RAG-enhanced systems will require unwavering commitment to interdisciplinary collaboration—across AI, domain sciences, ethics, policy, and user-centered design—as a crosscutting meta-principle. The next generation of AI-driven decision-support and discovery systems must be unequivocally user- and context-aware, seamlessly integrating robust retrieval, efficient and relevant augmentation, explainable interaction, and scalable automation. Achieving these outcomes demands sustained synthesis of expertise, transparency, and scientific rigor throughout the lifecycle of methods and applications.

References

[1] Maurice Abaho, Jialiang Guo, and Sebastien Harpe. 2024. Enhanced Dense Retrieval Knowledge Graph Augmentation. *Journal of Artificial Intelligence Research* 80 (2024), 1139–1178. https://jair.org/index.php/jair/article/view/14365

[2] J. Baek, A. Fikri Aji, and A. Saffari. 2023. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. *arXiv preprint arXiv:2306.04136* (2023). https://arxiv.org/abs/2306.04136

[3] J. Baek, N. Chandrasekaran, S. Cucerzan, A. Herring, and S. K. Jauhar. 2024. Knowledge-Augmented Large Language Models for Personalized Contextual Query Suggestion. In *Proceedings of The Web Conference (WWW) 2024.* https://arxiv.org/abs/2311.06318 arXiv preprint arXiv:2311.06318, to appear.

Table 23: Summary of Real-World Impacts, Recommended Practices, and Open Challenges in RAG-Enhanced Decision Support

Domain/Impact Area	Prominent Frameworks/Approaches	Recommended Practices	Open Challenges/Gaps
Biomedicine & Health	RAG-LLMs, Knowledge Graph Augmentation [17, 29, 39, 43]	Integration with guidelines, transparency, continual updating, explainable outputs	Data quality, regulatory compliance, bias, patient privacy, robustness to misinformation [17, 39, 57]
Legal & Public Policy	Topic embeddings, iterative retrieval, explainable LLMs [3, 4, 9, 50]	Provenance tracking, cross-jurisdictional adaptation, human-in-the-loop review	Scaling to large corpora, multilingual/cross-system consistency, interpretability vs. privacy [3]
Scientific Discovery	Multimodal/cross-domain RAG, KG integration [1, 47, 53, 64]	Workflow automation, literature mining, data-driven hypothesis generation, automated provenance	Representation of uncertainty, scalability, domain adaptation
Industrial/Recommendation Systems	Contrastive learning, context-aware augmentation [31, 37, 54, 56, 58]	Supervised augmentation, domain-specific fine-tuning, continuous model validation	Rare event detection, generalization across shifts, sample efficiency, explainability
All Domains	Modular RAG, transparent evaluation, active knowledge updating [3, 8, 16, 26, 34]	Stakeholder engagement, context-awareness, interdisciplinary synthesis	Balancing explainability and privacy, regulatory clarity, robust human-AI collaboration [3, 17, 45]

[4] C. Bai, X. Fan, J. Liu, W. Tang, H. Huang, and J. Yin. 2024. Graph Convolutional Network Discrete Hashing for Cross-Modal Retrieval. *IEEE Transactions on Neural Networks and Learning Systems* 35, 4 (2024), 1714–1727. https://ieeexplore.ieee.org/document/9779852

[5] X. Bai, S. He, Y. Li, Y. Xie, X. Zhang, W. Du, and J.-R. Li. 2025. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials* 11 (2025). doi:10.1038/s41524-025-01540-6

[6] O. Bergman, T. Israeli, and S. Whittaker. 2020. Factors hindering shared files retrieval. *Aslib Journal of Information Management* 72, 1 (2020), 130–147. doi:10.1108/AJIM-05-2019-0120

[7] O. Bergman and E. Shnaper-Reinberg. 2025. The effect of cooking recipe storage on their retrieval. *Journal of Documentation* ahead-of-print, ahead-of-print (2025). doi:10.1108/JD-01-2025-0031

[8] O. Bergman, S. Whittaker, and Y. Frishman. 2019. Let's get personal: the little nudge that improves document retrieval in the Cloud. *Journal of Documentation* 75, 2 (2019), 379–396. doi:10.1108/JD-06-2018-0098

[9] Federico Castagna, Sara Tonelli, and Serena Villata. 2024. Computational Argumentation-based Chatbots: a Survey. *Journal of Artificial Intelligence Research* 80 (2024), 1269–1330. doi:10.1613/jair.1.15407

[10] Tanmoy Chakraborty, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. 2023. Information retrieval algorithms and neural ranking models to detect previously fact-checked information. *Neurocomputing* 557 (2023), 126680. doi:10.1016/j.neucom.2023.126680

[11] H. Chen, Z. Chen, Y. Zhao, M. Wang, L. Li, M. Zhang, and M. Zhang. 2024. Retrieval-style In-Context Learning for Few-shot Hierarchical Text Classification. *Transactions of the Association for Computational Linguistics* 12 (2024). https://transacl.org/index.php/tacl/article/view/6137

[12] F. Dammak and H. Kammoun. 2021. Combining semi-supervised and active learning to rank algorithms: application to Document Retrieval. *Information Retrieval Journal* 24 (2021), 371–399. https://link.springer.com/article/10.1007/s10791-021-09403-7

[13] A. Dundar and I. Garcia-Dorado. 2017. Context Augmentation for Convolutional Neural Networks. *arXiv preprint arXiv:1712.01653* (2017). https://arxiv.org/abs/1712.01653

[14] C. Ehrett, S. Hegde, K. Andre, D. Liu, and T. Wilson. 2024. Leveraging Open-Source Large Language Models for Data Augmentation in Hospital Staff Surveys: Mixed Methods Study. *JMIR Medical Education* 10, 1 (2024), e51433. doi:10.2196/51433

[15] Xiyan Fu and Anette Frank. 2024. Exploring Continual Learning of Compositional Generalization in NLI. *Transactions of the Association for Computational Linguistics* 12 (2024), 912–932. doi:10.1162/tacl_a_00680

[16] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2024). https://arxiv.org/abs/2312.10997

[17] Stephen Gilbert, Jakob Nikolas Kather, and Aidan Hogan. 2024. Augmented non-hallucinating large language models as medical information curators. *npj Digital Medicine* 7 (2024), Article number: 100. doi:10.1038/s41746-024-01081-0

[18] T. C. Guetterman, T. Chang, M. DeJonckheere, T. Basu, E. Scruggs, and V. G. Vinod Vydiswaran. 2018. Augmenting Qualitative Text Analysis with Natural Language Processing: Methodological Study. *Journal of Medical Internet Research* 20, 6 (2018), e231. doi:10.2196/jmir.9702

[19] Zhipeng Gui, Xinjie Liu, Anqi Zhao, Yuhan Jiang, Zhipeng Ling, Xiaohui Hu, Fa Li, Zelong Yang, Huayi Wu, and Shuangming Zhao. 2024. Map retrieval intention recognition based on relevance feedback and geographic semantic guidance: For better understanding user retrieval demands. *Information Processing & Management* 61, 6 (2024), 103767. doi:10.1016/J.IPM.2024.103767

[20] Y. Guo, Q. Zhang, Z. Xie, and S. Jiang. 2024. Evaluating large language models for health-related text classification and question answering: A comparative study of domain-specific and general-purpose models. *Journal of the American Medical Informatics Association* 31, 10 (2024), 2181–2192. doi:10.1093/jamia/ocad243

[21] T. Gupta, M. Zaki, N. M. Anoop Krishnan, and Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials* 8 (2022), Article no. 102. https://www.nature.com/articles/s41524-022-00784-w

[22] M. Hindi, A. Smith, T. Chen, and P. Brown. 2025. Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 36, 2 (2025), 215–231. https://ieeexplore.ieee.org/iel8/6287639/10820123/10921633.pdf

[23] L. Huang, J. Yang, and Z. H. Zhang. 2022. A Comprehensive Review on Retrieval-Augmented Language Models. *IEEE Transactions on Neural Networks and Learning Systems* 33, 5 (2022), 2348–2361.

[24] T. K. W. Hung, G. J. Kuperman, E. J. Sherman, A. L. Ho, C. Weng, D. G. Pfister, and J. J. Mao. 2024. Performance of Retrieval-Augmented Large Language Models to Recommend Head and Neck Cancer Clinical Trials. *Journal of Medical Internet Research* 26, 1 (2024), e60695. https://www.jmir.org/2024/1/e60695

[25] K. Huseynova and J. Isbarov. 2024. Enhanced document retrieval with topic embeddings. *arXiv preprint arXiv:2408.10435* (Aug 2024). https://arxiv.org/abs/2408.10435

[26] G. Izacard, S. Touvron, F. Barbieri, A. Hosseini, N. Goyal, F. M. Sellam, K. Singh, E. Grave, T. Kocisky, E. J. M. Tromp, C. Lacroix, F. Raiss, F. Belinkov, N. Parikh, E. M. Khalifa, M. B. A. Haddad, A. Paria, N. H. E. Cesa-Bianchi, and S. Edunov. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research* 24, 68 (2023), 1–65. http://www.jmlr.org/papers/volume24/23-0037/23-0037.pdf

[27] H. Jaeger. 2017. Using Conceptors to Manage Neural Long-Term Memories for Temporal Patterns. *Journal of Machine Learning Research* 18, 13 (2017), 1–43. https://www.jmlr.org/papers/volume18/15-449/15-449.pdf

[28] M. Kang, J. M. Kwak, J. Baek, and S. J. Hwang. 2023. Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation. *arXiv preprint arXiv:2305.18846* (2023). https://arxiv.org/abs/2305.18846

[29] Y. H. Ke, L. Jin, K. Elangovan, H. R. Abdullah, N. Liu, A. T. H. Sia, C. R. Soh, J. Y. M. Tung, J. C. L. Ong, C.-F. Kuo, S.-C. Wu, V. P. Kovacheva, and D. S. W. Ting. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine* 8 (2025), Article no. 187. https://www.nature.com/articles/s41746-025-01519-z

[30] Julian Killingback, Hansi Zeng, and Hamed Zamani. 2025. Hypencoder: Hypernetworks for Information Retrieval. *arXiv preprint arXiv:2502.05364* (2025). https://arxiv.org/abs/2502.05364

[31] H. Kim, D. Kim, P. Ahn, S. Suh, H. Cho, and J. Kim. 2024. ContextMix: A context-aware data augmentation method for industrial visual inspection systems. *arXiv preprint arXiv:2401.10050* (2024). https://arxiv.org/abs/2401.10050 Accepted to EAAI.

[32] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. https://arxiv.org/abs/2005.11401 arXiv:2005.11401.

[33] Hai Li, Jingyi Huang, Mengmeng Ji, Yuyi Yang, and Ruopeng An. 2025. Use of Retrieval-Augmented Large Language Model for COVID-19 Fact-Checking: Development and Usability Study. *Journal of Medical Internet Research* 27 (2025). doi:10.2196/66098

[34] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized Prompt Learning for Explainable Recommendation. *ACM Transactions on Information Systems* 41, 4 (2023), 1–26. doi:10.1145/3524097

[35] Y. Li, J. Zhao, M. Li, Y. Dang, E. Yu, J. Li, Z. Sun, U. Hussein, J. Wen, A. M. Abdelhameed, J. Mai, S. Li, Y. Yu, X. Hu, D. Yang, J. Feng, Z. Li, J. He, W. Tao, T. Duan, Y. Lou, F. Li, and C. Tao. 2024. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. *Journal of the American Medical Informatics Association* 31, 9 (2024), 2030–2039. doi:10.1093/jamia/ocae129

[36] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2024. How Can Recommender Systems Benefit from Large Language Models: A Survey. *ACM Transactions on Information Systems* (2024). https://arxiv.org/abs/2306.05817

[37] Y.-T. Lin, A. Papangelis, S. Kim, S. Lee, D. Hazarika, M. Namazifar, D. Jin, Y. Liu, and D. Hakkani-Tur. 2023. Selective In-Context Data Augmentation for Intent Detection using Pointwise V-Information. *arXiv preprint arXiv:2302.05096* (2023). https://arxiv.org/abs/2302.05096 Accepted at EACL 2023.

[38] S. Liu, H. Chen, T. Wang, C. Zhang, Y. Wang, H. Wei, D. Wang, X. Yu, Y. Zhang, and M. Huang. 2025. A systematic review, meta-analysis, and clinical development of retrieval-augmented generation for large language model–enabled question answering in clinical practice. *Journal of the American Medical Informatics Association* 32, 4 (2025), 605–619. doi:10.1093/jamia/ocad348

[39] S. Liu, A. B. McCoy, and A. Wright. 2025. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association* 32, 4 (2025), 605–615. doi:10.1093/jamia/ocaf008

[40] X. Liu, Y. Wang, H. Wu, and L. Chen. 2025. RAG4DS: Retrieval-Augmented Generation for Data Spaces—A Unified Lifecycle, Challenges, and Opportunities. *IEEE Transactions on Neural Networks and Learning Systems* 36, 1 (2025), 77–92. https://ieeexplore.ieee.org/iel8/6287639/10820123/10902131.pdf

[41] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Robust Neural Information Retrieval: An Adversarial and Out-of-distribution Perspective. *arXiv preprint arXiv:2407.06992* (2024). https://arxiv.org/abs/2407.06992

[42] Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P. Ma, April S. Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, Nigam H. Shah, and Jonathan H. Chen. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine* 8 (2025). doi:10.1038/s41746-024-01377-1

[43] Chin Siang Ong, Nicholas T. Obey, Yanan Zheng, Arman Cohan, and Eric B. Schneider. 2024. SurgeryLLM: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. *npj Digital Medicine* 7 (2024), Article number: 364. doi:10.1038/s41746-024-01391-3

[44] Andrew Parry, Catherine Chen, Carsten Eickhoff, and Sean MacAvaney. 2025. MechIR: A Mechanistic Interpretability Framework for Information Retrieval. arXiv preprint arXiv:2501.10165. https://arxiv.org/abs/2501.10165 Demo paper, Proceedings of the European Conference on Information Retrieval (ECIR) 2025.

[45] V. L. Payne, U. Sattar, M. Wright, E. Hill, J. M. Butler, B. Macpherson, A. Jeppesen, G. Del Fiol, and K. Madaras-Kelly. 2024. Clinician perspectives on how situational context and augmented intelligence design features impact perceived usefulness of sepsis prediction scores embedded within a simulated electronic health record. *Journal of the American Medical Informatics Association* 31, 6 (2024), 1331–1340. doi:10.1093/jamia/ocae089

[46] Michael H. Prince, Henry Chan, Aikaterini Vriza, Tao Zhou, Varuni K. Sastry, Yanqi Luo, Matthew T. Dearing, Ross J. Harder, Rama K. Vasudevan, and Mathew J. Cherukara. 2024. Opportunities for retrieval and tool augmented large language models in scientific facilities. *npj Computational Materials* 10 (2024). doi:10.1038/s41524-024-01423-2

[47] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics* 11 (2023). https://transacl.org/index.php/tacl/article/view/5039

[48] Ohad Rubin and Jonathan Berant. 2024. Retrieval-Pretrained Transformer: Long-range Language Modeling with Self-retrieval. *Transactions of the Association for Computational Linguistics* 12 (2024). https://transacl.org/index.php/tacl/article/view/6313

[49] M. Solanki. 2025. Efficient Document Retrieval with G-Retriever. *arXiv preprint arXiv:2504.14955* (April 2025). https://arxiv.org/abs/2504.14955

[50] P. Staszewski, M. Jaworski, J. Cao, and L. Rutkowski. 2022. A New Approach to Descriptors Generation for Image Retrieval by Analyzing Activations of Deep Neural Network Layers. *IEEE Transactions on Neural Networks and Learning Systems* 33, 7 (2022), 3075–3083. https://ieeexplore.ieee.org/document/9451541

[51] M. Trabelsi, Z. Chen, B. D. Davison, and J. Heflin. 2021. Neural ranking models for document retrieval. *Information Retrieval Journal* 24 (2021), 400–444. https://link.springer.com/article/10.1007/s10791-021-09398-0

[52] R. Upadhyay and M. Viviani. 2025. Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy. *Information Retrieval Journal (now Discover Computing)* 28, Article 27 (2025), 44 pages. https://link.springer.com/article/10.1007/s10791-025-09505-5

[53] Benigno Uria, Iain Murray, Stephan Ren, Risto Piché, Aaron Courville, and Hugo Larochelle. 2016. Neural Autoregressive Distribution Estimation. *Journal of Machine Learning Research* 17, 205 (2016), 1–37. https://www.jmlr.org/papers/volume17/16-272/16-272.pdf

[54] Chenyang Wang, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. Sequential Recommendation with Multiple Contrast Signals. *ACM Transactions on Information Systems* 41, 1 (2023), 1–27. doi:10.1145/3522673

[55] D. Wang, J. Liang, J. Ye, J. Li, J. Li, Q. Zhang, Q. Hu, C. Pan, D. Wang, Z. Liu, W. Shi, D. Shi, F. Li, B. Qu, and Y. Zheng. 2024. Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study. *Journal of Medical Internet Research* 26, 1 (2024), e58041. https://www.jmir.org/2024/1/e58041/

[56] Dong Wang, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. Large-Scale Pre-Training for Sequential Recommendation with Contrastive Learning. *ACM Transactions on Information Systems* 41, 2 (2023), 1–23. doi:10.1145/3570620

[57] H. Wang, A. Prasad, E. Stengel-Eskin, and M. Bansal. 2025. Retrieval-Augmented Generation with Conflicting Evidence. *arXiv preprint arXiv:2504.13079* (2025). https://arxiv.org/abs/2504.13079

[58] Chuyuan Wei, Ke Duan, Shengda Zhuo, Hongchun Wang, Shuqiang Huang, and Jie Liu. 2025. Enhanced Recommendation Systems with Retrieval-Augmented Large Language Model. *Journal of Artificial Intelligence Research* 82 (2025), 1–27. https://jair.org/index.php/jair/article/view/17809

[59] Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Computational Linguistics* 45, 1 (2019), 163–197. doi:10.1162/coli_a_00345

[60] Lanling Xu, Junjie Zhang, Bingqian Li, Jinpeng Wang, Sheng Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2025. Tapping the Potential of Large Language Models as Recommender Systems: A Comprehensive Framework and Empirical Analysis. *ACM Transactions on Information Systems* (2025). https://arxiv.org/abs/2401.04997

[61] T. Yang, M. Song, Z. Zhang, H. Huang, W. Deng, F. Sun, and Q. Zhang. 2023. Auto Search Indexer for End-to-End Document Retrieval. *arXiv preprint arXiv:2310.12455* (Oct. 2023). https://arxiv.org/abs/2310.12455

[62] Z. Zhan, S. Zhou, M. Li, and R. Zhang. 2025. RAMIE: retrieval-augmented multi-task information extraction with large language models on dietary supplements. *Journal of the American Medical Informatics Association* 32, 3 (2025), 545–554. doi:10.1093/jamia/ocaf002

[63] Gongbo Zhang, Zihan Xu, Qiao Jin, Fangyi Chen, Yilu Fang, Yi Liu, Justin F. Rousseau, Ziyang Xu, Zhiyong Lu, Chunhua Weng, and Yifan Peng. 2025. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine* 8 (2025). doi:10.1038/s41746-025-01651-w

[64] L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou. 2022. Deep Multimodal Transfer Learning for Cross-Modal Retrieval. *IEEE Transactions on Neural Networks and Learning Systems* 33, 2 (2022), 798–810. doi:10.1109/TNNLS.2020.3032604