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, 18 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

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). These advances bear transformative implications, especially for high-stakes, knowledge-intensive domains such as biomedicine, healthcare, and law. In contrast to traditional language models, which predominantly rely on static, parametric knowledge encoded during pre-training, RAG frameworks integrate large language model (LLM) architectures with external retrieval mechanisms. This fusion equips LLMs to access, incorporate, and reason over dynamic, domain-specific sources, thereby addressing the limitations of static knowledge and significantly enhancing accuracy, transparency, and adaptability in 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. Here, stringent requirements for transparency, explainability, and scalable deployment converge with domain-specific challenges. In medicine, RAG-based systems have consistently outperformed their non-augmented counterparts across diverse tasks, including clinical decision support, guideline adherence, and the detection of misinformation. These systems deliver improved factual accuracy and foster greater user trust [5, 16, 17, 20–23, 29, 33, 35, 38, 39, 42, 43, 52, 55, 62]. Legal applications, similarly, leverage RAG pipelines to enhance knowledge provenance, support regulatory compliance, and ensure procedural integrity through verified and transparent retrieval [11, 16, 23, 32, 39, 48].

Notwithstanding these advances, RAG and knowledge-augmented models face notable limitations. A persistent challenge is hallucination—the generation of plausible yet unsupported content—which carries amplified risks in settings where model errors can undermine patient safety, legal accountability, or public confidence [19, 25, 33, 38, 40, 45, 47, 49, 55, 61]. Further, these models are often hampered by outdated or incomplete knowledge bases, and their robustness to out-of-distribution (OOD) data remains insufficiently validated in real-world deployments. Mission-critical scenarios further necessitate reliable handling of privacy constraints, compliance with evolving regulatory requirements, scalable operation in complex and multi-turn interactions, and explicit management of biases that may be inherited or exacerbated by both retrieval and generation modules [19, 25, 33, 38, 40, 45, 47, 49, 55, 61].

Collectively, these challenges reveal a central paradox: while RAG and related technologies hold promise for improved factuality, adaptability, and user trust, their operationalization introduces new vectors for error, instability, and bias. The literature increasingly points to the necessity of continual model updates, rigorous and transparent benchmarking, and robust provenance

tracking [33, 47, 55]. There is a parallel movement toward integrating structured external resources, such as knowledge graphs, to buttress the statistical strengths of LLMs with verifiable and regulatable knowledge bases [11, 49].

1.2 Scope and Contributions

This survey aims to unify and critically examine the foundational techniques, systems architectures, and evaluation methodologies underlying retrieval- and context-augmented generation. Coverage extends across the full RAG implementation stack:

- Classical and neural information retrieval approaches (sparse, dense, hybrid)
- Strategies for data and context augmentation
- Contrastive learning paradigms
- Knowledge graph construction and integration
- Collection of architectural variants within RAG
- Evaluation frameworks tailored to retrieval-augmented and generative systems

By systematically analyzing these components, this review clarifies the interplay of recent advances that collectively drive performance and reliability in knowledge-intensive applications.

A distinguishing feature of this survey is its multidomain perspective, focusing on biomedical, legal, and general-purpose settings, with particular emphasis on applications involving vision or intent detection [5, 13, 16, 17, 20–24, 29, 33, 35, 37–40, 42, 43, 46, 52, 55, 62]. This review systematically maps the landscape of core RAG use cases, such as:

- Clinical question answering
- Clinical and legal decision support
- Misinformation mitigation
- Recommender systems
- Legal workflow and pipeline optimization
- Intent detection with multimodal signals

This comprehensive mapping elucidates both the diversity of RAG deployments and the distinct technical and regulatory considerations each domain imposes [5, 16, 17, 20–22, 24, 29, 33, 35, 38, 39, 42, 43, 46, 52, 55, 62].

This survey places special emphasis on approaches that move beyond surface-level augmentation, focusing instead on frameworks that embed RAG into robust, scalable, and interpretable machine reasoning systems. Topics addressed include advanced retrieval methods (entity-based, knowledge graph-driven, multimodal), techniques in representation learning, strategies for model grounding, and regulatory-aware workflow design. Moreover, critical assessment highlights unresolved issues relating to robustness, fairness, regulatory and privacy considerations, and deployment scalability, while also charting possible directions for future research and standardization.

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.

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

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 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—provide robust baselines for query-document matching by relying on sparse lexical or frequency-based interactions [12, 38, 52]. These models are effective in structured corpora where token-level matching suffices. However, as data complexity and domain heterogeneity increase, especially in scientific, clinical, and legal collections, these approaches reveal substantial shortcomings in handling synonymy, semantic drift, and expressive matching [12, 38, 51].

The advent of neural and dense retrieval paradigms addresses these limitations by encoding queries and documents as dense vectors, enabling more sophisticated semantic matching. Notable architectures include bi-encoders, dual-encoders, and advanced frameworks such as Hypencoder [3, 11, 16, 19, 23, 26, 28, 30, 32, 33, 38, 39, 41, 47, 48, 51, 55, 59]. Hybrid models, combining traditional term-based and neural dense retrieval approaches, have demonstrated strong performance, especially in Retrieval-Augmented Generation (RAG) pipelines and other knowledge-intensive systems [33, 38, 51, 52].

Further refinement is achieved via interaction-based neural ranking models, which model the complex interplay between queries and documents by leveraging attention mechanisms and contextual embeddings for fine-grained relevance estimation [3, 11, 16, 19, 26, 28, 30, 32, 33, 38, 39, 41, 47, 48, 51, 55, 59]. For instance, sequential matching models address conversational context and utterance relationships, outperforming simple vector-based retrieval in dialogue applications [51]. Nevertheless, the expressiveness of these models comes at the expense of increased computational requirements, presenting challenges in scalability and long-context management [33, 55].

Personalization represents a critical frontier in IR development. Entity-centric knowledge bases and context-aware augmentation enable retrieval systems to deliver tailored results based on user history and domain-specific criteria, enhancing both recommendation quality and contextual relevance [16, 19, 30, 32, 33, 38, 41, 47, 48, 51, 59].

Despite significant progress, neural IR models remain susceptible to adversarial or out-of-distribution queries and are sensitive to domain shifts [4, 22, 25, 33, 40, 62]. Major research directions focus on:

- **Robustness**: Addressing performance degradation under challenging input conditions.
- **Interpretability**: Illuminating neural model decision processes.
- **Benchmarking**: Standardizing evaluations with heterogeneous datasets, as exemplified by the BestIR suite [22, 40, 62].

Developing harmonized definitions of robustness and implementing defenses for neural retrieval remain open challenges, particularly as integration with LLMs introduces added complexity [22, 40, 62].

2.3 Knowledge and Context Augmentation

Effective handling of domain-specific knowledge gaps and reliable inference in LLMs increasingly depends on knowledge and context augmentation. The following strategies play a pivotal role in modern RAG workflows and data-centric AI:

- **Query Expansion and Synthetic Data Generation**: Techniques such as mixup, chunking, and prompt engineering generate diverse training scenarios to overcome annotation scarcity [1, 3, 11, 13, 17, 21, 26, 28, 32, 33, 37–39, 47, 51, 53, 55, 60, 63].
- **Teacher-Student Knowledge Distillation**: Structured transfer of competencies to smaller or domain-adapted models, improving robustness and data efficiency [36, 55, 60].
- **Active Learning and Feedback**: Iterative model refinement using pseudo-labeling and selective human annotation [21, 55, 60].
- **Chunking and Context Selection**: Pipelines such as CLEAR facilitate accuracy and efficiency in entity extraction and biomedical NLP [3, 33, 37, 39, 51].

Integration with knowledge graphs and knowledge-grounded neural architectures constitutes a particularly transformative advance:

- LLMs may directly interact with, augment, or be augmented by structured representations—enabling verifiable outputs, enforcing factual consistency, and supporting multi-hop reasoning [1, 5, 16, 17, 21, 22, 28, 32, 33, 38, 42, 49].
- Knowledge graph injection, as employed in scientific, biomedical, and legal applications, yields superior representational fidelity and equips models to handle rare entities, mitigate hallucinations, and support compliance with regulatory requirements for verifiable AI [5, 22, 33, 42, 49].

Diverse augmentation strategies allow practitioners to tailor solutions according to operational requirements, balancing computational cost with responsiveness and knowledge richness [1, 13, 17, 32, 33, 37, 55]. The contemporary trend toward modular, hybrid architectures—featuring pluggable augmentation modules—enables advances in explainability, adaptation, privacy, and scalability [21, 36, 55, 60].

As shown in Table 1, these diverse techniques collectively underpin advances in robust, domain-aligned, and verifiable AI.

2.4 In-Context Data Augmentation Techniques

As LLMs and vision models proliferate in domains characterized by limited labeled data and stringent regulatory demands, in-context data augmentation has become indispensable. Advanced methods synergize pretrained language models with pointwise information metrics (such as V-information), intent-sensitive filtering, and synthetic data generation to enhance sample efficiency, particularly in intent detection and hierarchical text classification tasks [37]. Selectively incorporating augmented samples—based on their marginal utility—yields state-of-the-art performance while mitigating overfitting and noise [37].

Vision domains benefit similarly from innovative approaches such as dynamic segmentation and controlled background–foreground combinations during data synthesis, delivering particular strengths

Table 1: 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

under limited or synthetic data regimes [13]. These findings emphasize the necessity for alignment between augmentation strategies, model architecture, and statistical data properties.

A salient application is the use of open-source LLMs (for example, LLaMA and Alpaca) in the synthetic augmentation of hospital survey datasets [14]. Deploying models locally preserves privacy and cost efficiency while expanding training corpora in sensitive clinical environments where access to authentic narratives is restricted. The integration of high-quality synthetic samples has been empirically demonstrated to robustly improve classifier accuracy, validating the viability of LLM-driven augmentation under data scarcity and privacy constraints [14].

Overall, the evolution of data augmentation techniques encompasses a broad spectrum:

- **Intelligent Prompt Engineering**: Crafting prompts to generate diverse, relevant synthetic data.
- **Intent-Aware Sample Selection**: Filtering augmented data by utility or informativeness.
- **Domain-Adapted Synthetic Generation**: Tailoring data to match desired statistical and operational domain properties.

The rigorous integration of these approaches into retrieval-augmented models and domain adaptation frameworks holds the key to developing robust, transparent, and high-performing AI systems across scientific, clinical, legal, and multimodal contexts [1, 3, 11, 13, 14, 17, 21, 26, 28, 32, 33, 37–39, 47, 51, 53, 55, 60, 63].

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 followed a trajectory from monolithic to modular, interoperable designs supporting scalable deployment and advanced knowledge integration. High-level RAG data space models (RAG-DSMs) systematize the RAG workflow within federated, secure, and interoperable data infrastructures, supporting cross-institutional knowledge exchange and fostering trust—especially in regulated domains [40].

Central to these advancements are modular retriever-generator pipelines, now capable of integrating feedback mechanisms whereby generation quality iteratively refines future retrievals and vice versa [3, 19, 22, 23, 26, 28, 30, 33, 36, 39, 40, 47–49]. Innovations in document identifier (docid) management—including direct docid generation and generative retrieval models—expand the capacity for dynamic, scalable retrieval as knowledge resources evolve [33, 38, 61]. Additionally, cognitive information retrieval (IR) pipelines, which

merge symbolic reasoning with neural methods, offer greater interpretability without sacrificing the expressive power of deep learning models [21, 28, 49].

The integration with distributed data spaces is a landmark feature, enabling secure data sharing and collaboration among trusted parties in sensitive environments [40]. Such systems support organizational interoperability, compliance with legal frameworks (e.g., GDPR, HIPAA), real-time updating, and robust auditing—all while maintaining scalability and low-latency performance.

A high-level comparison of select architectural innovations is presented in Table 2.

3.3 Advanced Retrieval and Context Management

As RAG models advance, the sophistication of retrieval methods has become pivotal to performance and adaptability. Hybrid retrieval architectures that jointly leverage sparse and dense signals, as well as enriched, graph-based retrieval, have outperformed traditional approaches in retrieval accuracy and domain robustness [1, 5, 16, 17, 21, 28, 32, 33, 38, 42, 49]. Techniques such as selective subgraph construction, guided by task-specific attention, have further improved efficiency by narrowing retrieval to contextually relevant knowledge units [17, 38].

Recently proposed paradigms—including logic-of-task (LOT) retrievers, agentic approaches such as agentic/LOT-RAG, CRAG, and SRAG, and contextually adaptive retrieval methodologies—enable the dynamic configuration of retrieval based on user workflows and evidentiary needs [33, 40]. These agent-driven designs support optimized interactions between retrieval and generation, particularly crucial in applications where transparency and dynamic augmentation are imperative, such as clinical question answering and pandemic-related fact verification [33, 40].

Efficient context management remains a challenge, especially for domains that require processing lengthy, unstructured documents with intricate dependencies. To address limitations such as context window overflow and the "lost-in-the-middle" effect, several techniques have demonstrated efficacy:

- **Input Segmentation:** Dividing documents into semantically coherent chunks to maximize context retention [11, 32, 33, 38, 39, 47, 51, 55, 63].
- **Map-Reduce Partitioning:** Efficiently processing subdocuments in parallel for scalable generation.
- **Dynamic Context Prioritization:** Selecting or re-ordering context windows to ensure the inclusion of salient information while minimizing token usage.

Map-reduce-based RAG variants and advanced context prioritization strategies have reduced computational load while preserving extraction accuracy, particularly in demanding settings such as electronic health record (EHR) pipelines [38, 39, 55].

Emerging best practices emphasize the design of domain-driven RAG pipelines, treating data provenance, security, and transparency as integral system metrics [22, 40]. Iterative development frameworks structure deployment around distinct pre-retrieval, retrieval, and post-retrieval cycles, allowing for agile adaptation to regulatory shifts and ongoing improvement [22, 39].

In summary, the trajectory of RAG research is characterized by the emergence of deeply integrated, contextually adaptive, and trustworthy systems. These advances couple state-of-the-art retrieval techniques with robust generation architectures, underpinning the scalable and transparent deployment of LLMs across the most knowledge-intensive and high-stakes domains.

4 Contextual Data Augmentation, Contrastive Learning, and Multimodal Applications

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

Table 2: 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

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

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, if selection and filtering are inadequate, generated content may introduce semantic drift or noise, ultimately impairing model performance. Recent advancements address this by leveraging pointwise V-information (PVI) to quantify the utility of each synthesized sample, admitting only high-value augmentations into the training corpus. This results in state-of-the-art accuracy in both few-shot and full-shot scenarios [37]. The findings underscore the necessity of stringent calibration and quality control during generative augmentation, particularly for low-resource, intent-driven applications.

Augmentation strategies in the visual domain have similarly evolved, expanding beyond conventional pixel-level manipulations. Approaches that blend background variation with foreground segmentation have shown clear benefits, especially in settings with sparse or imbalanced data [13]. The ContextMix technique exemplifies these advances by combining resized, context-rich image regions, thereby producing more discriminative and context-aware examples. By harmonizing object and environmental cues, ContextMix not only enhances classification and detection accuracy but also bolsters robustness against adversarial perturbations and class imbalance. This is especially advantageous in industrial defect detection domains characterized by limited, imbalanced datasets. Furthermore, the method's minimal computational overhead and straightforward formulation support its applicability in practical manufacturing environments [31].

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 each modality's statistical and semantic properties. Transfer learning techniques—such as deep multimodal transfer and pseudo-labeling—help propagate knowledge from richly annotated source domains to underrepresented target domains, even when label sets differ [3–5, 7, 13, 17, 19, 21, 27, 28, 30, 33, 36, 37, 46–48, 51, 53, 55, 61, 64]. Nevertheless, challenges remain:

- Preserving semantic alignment across modalities, particularly in the presence of modality asymmetry.
- Ensuring consistent quality and relevance of generated augmentations.
- Addressing high intra-class variance and avoiding training instability in low-resource circumstances.

Despite progress, several open problems persist. Synthesized or contextually mixed samples can mislead models if contextual or object boundaries are not appropriately maintained. Furthermore, variability in augmentation quality may introduce bias or reduce model stability, highlighting the need for more adaptive, quality-assured 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 [38].

However, the transition to real-time adaptation poses significant challenges:

- Managing evolving, non-stationary user preferences.

- Maintaining user privacy and compliance with regulatory frameworks.
- Scaling adaptive personalization to diverse platforms and linguistic environments.

There is now broad agreement that effective adaptive context modeling requires joint optimization for transparency, fairness, and privacy. This underscores the increasing relevance of federated and on-device learning, privacy-preserving embeddings, and interpretable user modeling frameworks as future research directions.

4.4 Synthesis and Open Challenges

In sum, the convergence of contextual data augmentation and contrastive learning—deployed at both the data and model levels—has proved instrumental in addressing the requirements of multimodal, low-resource, and personalized information retrieval and recommendation contexts. Nevertheless, several critical challenges remain:

- Harmonizing data augmentation techniques with the intricate demands of adaptive user modeling.
- Scaling contrastive learning to high-dimensional, sparse, or heterogeneous input spaces.
- Systematically evaluating the ethical and privacy implications associated with increasingly personalized and context-aware systems.

Ongoing advances in augmentation pipelines, cross-modal signal alignment, and privacy-centric modeling will be crucial to realizing robust, fair, and scalable IR and recommendation systems.

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

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, 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].

As summarized in Table 3, while RAG pipelines have markedly improved accuracy and transparency in clinical, biomedical, and public health domains, ongoing challenges in data quality, update latency, interpretability, and privacy remain important areas for future research and operational refinement.

5.2 Legal, Regulatory, and Security Applications

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

Particular emphasis is placed on privacy-preserving data architectures. Compliant retrieval mechanisms, such as federated or decentralized data handling, help guarantee that sensitive client or patient information remains 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]. Research foregrounds the imperative for rigorous risk management in conjunction with practical functionality; this entails integrating risk-aware retrieval strategies, policy-constrained generation modules, and traceable attribution of knowledge sources to withstand adversarial scrutiny and comply with legal discovery requirements [3–5, 7–11, 15, 16, 19, 22, 25–29, 32–34, 36, 39, 41, 42, 45–48, 50, 51, 53, 55, 59–61, 63, 64].

A significant requirement in legal decision support is explainability. Legal professionals necessitate not only direct answers but also actionable rationales that are firmly anchored in statutory law, caselaw, and procedural precedents. Retrieval-augmented systems enable traceable chains of reasoning and counterfactual analysis, establishing a robust foundation for future explainable legal AI systems that can meet both regulatory and societal expectations [22].

However, several open research challenges remain:

- **Cross-jurisdictional Scalability:** Adapting RAG pipelines to handle multi-jurisdictional and cross-lingual legal scenarios.
- **Transparency vs. Efficiency:** Balancing workflow transparency with the efficiency demands of legal practice.
- **Explainability:** Enhancing the interpretability and auditability of AI-generated legal outputs.

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. When these augmentations are incorporated into multimodal RAG architectures, they enrich contextual retrieval for downstream tasks by providing diverse, information-rich representations [13].

Increasingly, modern pipelines enable multimodal and cross-lingual retrieval, allowing for seamless integration and joint reasoning across text, image, graph, and tabular data. Key enabling technologies include deep multimodal transfer learning, cross-modal hashing enhanced by graph convolutional networks, and the deployment of optimized index/search strategies for retrieval in complex scientific and legal domains lacking exhaustive labeled data [13, 37, 47, 48]. This capability is particularly crucial in domains where evidence extends across documents, figures, and structured databases, supporting advanced vision-language models that facilitate document analysis, benchmarking, and multidisciplinary workflows [3, 7, 8, 19, 28, 30, 33, 36, 39, 48, 51, 55, 61].

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

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

Table 3: 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

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 4, 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 task-representative, high-quality datasets spanning diverse domains and modalities. In biomedical and clinical research, curated corpora such as PubMed, MIMIC, UMLS, BioASQ, MedQA-US, and MedMCQA enable evaluation on knowledge-intensive and reasoning tasks. In contrast, resources like Twitter and OpenDialKG extend assessment to social media and open-domain conversational contexts [1–5, 7–11, 13, 15, 16, 20, 21, 25–30, 32–39, 41, 42, 46–50, 52, 53, 55, 57–60, 62–64].

Synthetic datasets, constructed for purposes such as continual compositional inference and adversarial OOD testing, have become essential components of robust model evaluation [13, 33, 37, 55]. However, the strengths and limitations of each dataset type must be critically considered:

- **Annotated benchmarks** in domains such as clinical or legal offer structured and interpretable evaluation, yet often encounter challenges regarding scale, dynamic gold standards, and coverage bias.
- **Open-domain and vision-centric datasets** (e.g., IMAGENET1M, MatSci) support broader generalization assessments but may lack annotation granularity.
- **Multilingual and multimodal datasets** remain scarce, especially those with high-quality, gold-standard annotations, severely constraining progress in low-resource and cross-domain generalization.

Advances in knowledge graph extraction, domain adaptation (for example, MatSciBERT and KG-FM in materials science), and multi-modal synthesis—integrating vision and language modalities—underline the maturation of benchmarks beyond text-centric paradigms [2–4, 28, 36, 37, 49, 58, 62, 64]. Nonetheless, the persistent lack of naturalistic, user-generated queries paired with corresponding annotated gold answers—particularly in languages other than English—continues to hamper end-to-end benchmarking. This gap highlights an urgent need for collaborative dataset curation and standardization efforts to advance evaluation rigor and inclusivity.

Table 4: 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

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

Interpretability, security, and human oversight are increasingly central dimensions in the evaluation and deployment of RAG systems as they transition into mission-critical roles. Evaluation strategies are shifting toward user- and context-centered risk audits, with a focus on transparency and causal traceability of outputs—especially in settings where model decisions directly affect clinical, scientific, or legal outcomes [3, 5–10, 16, 17, 22, 24, 26, 27, 32–34, 36, 38–42, 44, 46, 55, 59, 60].

Explainability requirements now extend beyond retrospective justifications, demanding the capacity for prospective rationales that enable trust, troubleshooting, and compliance with regulatory frameworks [3, 6–8, 22, 33, 34, 36, 38, 40, 55, 60]. Early adoption of causal interpretability frameworks has begun to shed light on the diagnostic aspects of deep IR and RAG systems, attributing predictions or errors to specific model components or data features, thereby informing continual system improvement and targeted debugging [7, 17, 22, 33, 40, 46, 55].

Comparative evaluation protocols—blending human and LLM-based annotation—facilitate efficient, large-scale benchmarking, yet highlight the enduring necessity for human adjudication in instances of subjective or context-dependent model outputs [6–8, 20, 24, 33, 45, 55].

Security concerns have also become prominent as RAG systems are entrusted with sensitive data across healthcare, legal, and open-domain settings. The imperatives of privacy-preserving computation, trustworthy data sharing, and regulatory alignment have led to technical innovations such as the integration of RAG with secure data spaces, federated learning, and granular access controls [22, 40]. Striking a balance between data utility and privacy—particularly across organizational or jurisdictional boundaries—remains a key technical and legal challenge.

Finally, the integration of human-in-the-loop paradigms is proving indispensable to both evaluation and operational reliability. The incorporation of domain experts into risk audit processes ensures that AI-generated recommendations or extractions undergo rigorous contextual scrutiny, illuminating latent failure modes, informing user trust calibration, and providing feedback essential for system refinement and continuous learning [3, 5–10, 16, 17, 22, 24, 26, 27, 29, 32–34, 36, 38–42, 44, 46, 55, 59, 60].

Through the synthesis of advanced evaluation protocols, benchmark datasets, and interpretability frameworks, the field continues to navigate the intertwined challenges of reliability, transparency, and ethical deployment in RAG-driven AI. Persistent open problems underscore the importance of interdisciplinary collaboration—integrating technical, human factors, and policy expertise—to achieve trustworthy and robust AI systems.

7 Robustness, Ethics, Responsible Deployment, and Workflow Integration

7.1 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 underscores significant progress in mitigating vulnerabilities using retrieval-augmented generation (RAG) approaches, domain-adaptive indexing, and more robust neural architectures. For example, survey evidence indicates that despite technical advances, state-of-the-art dense and hybrid retrieval models remain susceptible to sophisticated adversarial attacks and OOD conditions. Dynamic adaptation strategies and continual learning paradigms are increasingly recognized as essential defenses against such challenges, although their application remains relatively underexplored [62].

The field has responded with technological innovations—including dynamic chunking, context prioritization, and multi-agent debate protocols—that have achieved demonstrable gains in reducing hallucinations, lowering misinformation dissemination, and enhancing the reliability of algorithmic recommendations. These benefits have been observed in diverse applications, ranging from perioperative medical guidance to automated fact-checking and legal analyses [3, 4, 10, 13, 22, 24, 28, 32, 33, 37, 38, 40, 41, 57, 63]. Despite these advancements, persistent challenges arise, particularly at the interface between system-level design and domain-specific knowledge integration. While adversarial robustness is typically evaluated in isolation, operational deployments often confront overlapping threats—such as conflicting evidence, ambiguity, and misinformation—necessitating simultaneous multi-faceted defenses.

The introduction of new datasets (e.g., RAMDocs) and frameworks (such as MADAM-RAG) has facilitated comprehensive error analysis, illuminating the limitations of existing RAG and LLM systems when exposed to compounded adversarial conditions [13]. Mechanistic strategies that combine dynamic retrieval, debate-oriented model architectures, and topic-enhanced embeddings have proven especially beneficial for both output stabilization and systematic failure mode analysis [22, 40]. However, the ongoing challenges of domain-specific variability, rapid corpus expansion, and model interpretability impede the full realization of robust OOD generalization and transparent error management [24, 33].

7.2 Ethical, Privacy, and Regulatory Considerations

Beyond technical robustness, ethical and legal accountability represent foundational pillars for deploying advanced retrieval and generative models. Key ethical issues encompass:

- Data-driven disparities and annotation bias
- Fairness in algorithmic recommendations and support decisions
- Requirements for privacy preservation and regulatory compliance (particularly in healthcare and law)

Addressing annotation and data biases is particularly critical, as recent studies demonstrate these can exacerbate inequities for marginalized or underrepresented populations, influencing both the fairness of models and the equity of their outputs [29, 33, 42].

In healthcare, evolving RAG frameworks and iterative error management schemes have been developed to ensure traceability and privacy by integrating both local and external data sources—while maintaining adherence to regulatory standards such as GDPR and HIPAA [47, 48, 53]. The preservation of privacy in LLM and RAG systems is especially challenging because performance improvements often require access to sensitive or proprietary information. To safeguard such data, leading approaches prioritize federated retrieval, robust access controls, and privacy-preserving user embeddings [50, 55]. Furthermore, several urgent research directions have emerged:

- Harmonization of regulatory requirements across jurisdictions
- Automation of regulatory compliance verification
- Enhancement of explainability and auditability, especially in scenarios with cross-border implications [5, 46, 59]

7.3 Interpretability and Human Collaboration

The inherent opacity of neural models, particularly in high-stakes environments, necessitates a robust commitment to interpretability, explainability, and human-in-the-loop (HITL) validation mechanisms. Mechanistic interpretability aims to correlate internal model computations with observable decisions, facilitating both causal understanding and targeted interventions [6–9, 22, 27, 33, 34, 36, 40, 44, 45, 55, 59, 60]. Despite technological advances, practitioners—including clinicians, legal experts, and end-users—report persistent discomfort related to the "black box" nature of LLMs, and often require direct access to model provenance, contributory evidence, and validation assets [8, 33, 55].

Modern deployment strategies increasingly incorporate techniques such as chain-of-thought prompting, computational argumentation frameworks, and counterfactual visualization, all of which foster transparency and improve user comprehension [6, 7, 9, 22, 34, 40, 44]. The integration of argumentation engines in LLM-driven chatbots and decision aids has been shown to enhance both transparency and the perceived trustworthiness of such tools in legal and healthcare settings [7, 8, 33]. Notably, most leading LLMs do not yet provide robust, built-in reasoning explainability, thereby identifying a critical need for hybrid systems that merge LLM fluency with structured modular reasoning.

Collaborative workflows that incorporate domain experts—through HITL validation—are central to resolving edge cases, verifying contextual accuracy, and progressively refining model outputs [22, 33, 40, 45].

7.4 User Interfaces and Workflow Integration

The effectiveness of robust and ethical AI systems ultimately depends upon the quality of user interfaces and their seamless integration into professional workflows. Recent studies underscore that environments such as clinics and legal practices require more than just transparent recommendations: interfaces must be behaviorally attuned, collaboration-enabling, and compatible with established documentation and triage routines [6–8, 22, 24, 33, 38, 40, 45, 55]. Successful deployments typically leverage:

- Decision-support dashboards
- Provenance-aware evidence visualizations
- Interactive feedback loops for human oversight and intervention

For example, early warning systems integrated within electronic health records (EHRs) and interactive document categorization platforms are shown to not only increase user satisfaction but also improve retrieval speed, accuracy, and overall trust in the AI system [6, 24, 33, 55].

In collaborative settings, the incorporation of AI-generated recommendations introduces further complexity, necessitating sophisticated solutions for versioning, access control, and transparency in shared documentation environments [24, 33, 38]. Empirical findings advocate for active user involvement in functions such as document categorization and retrieval augmentation, as opposed to passive automation. This approach leads to superior retrieval performance and enhanced knowledge retention [33, 55]. Interfaces that facilitate such engagement—while minimizing cognitive load and providing actionable explanations—are increasingly understood as essential for the responsible integration of AI into mission-critical domains [22, 40, 45].

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 and illuminate key methodological trends shaping current and future directions in the field.

8.1 Continual and Sequential Learning

Continual and sequential learning methodologies equip AI systems with the ability to adapt to dynamic domains, evolving tasks, and shifting user requirements over extended periods, while minimizing catastrophic forgetting and maintaining performance on previously

learned tasks. Notable research has foregrounded a diverse spectrum of techniques, including lifelong adaptation, hierarchical domain/task learning, knowledge transfer mechanisms, strategic data augmentation, and modular architectures for ongoing knowledge integration [4, 5, 13, 15, 25–27, 33, 37, 40, 41, 46, 50, 53, 55, 62, 64].

A prominent example is the CLEAR system, which operationalizes continual adaptation in the clinical domain. By integrating dynamic clinical named entity recognition with modular information retrieval, CLEAR achieves efficient inference and robust, scalable extraction of emergent knowledge as clinical documentation practices and standards evolve [62]. Its validation on longitudinal electronic health record (EHR) datasets demonstrates that explicit task- and domain-specific module incorporation can yield rapid generalization and effective transfer amid shifting real-world distributions.

Complementary to such engineering solutions, the C2Gen NLI challenge serves as a benchmark for the compositional continual generalization abilities of neural models. Results show that most models face significant difficulties in transferring primitive inference knowledge across sequentially ordered tasks; only dependency-aware, explicitly structured curricula succeed in reducing catastrophic forgetting [25]. These findings emphasize the need for explicit continual learning strategies, particularly in systems expected to dynamically accumulate and recombine knowledge.

Recent surveys addressing robustness in neural information retrieval underscore enduring challenges in out-of-distribution (OOD) adaptation, adversarial resilience, and the practical implementation of defenses against evolving attacks [33]. The paucity of unified OOD benchmarks and real-world deployment protocols leaves a gap in continual adaptation solutions for deployed systems. This motivates the generation of synthetic adversarial/OOD examples via LLMs and the instantiation of robust, adaptive evaluation pipelines.

In the context of scientific discovery and automated experiment platforms, context-aware LLMs such as CALMS perform dynamic retrieval, tool orchestration, and conversational memory management across scientific instrument operations, further demonstrating the value of lifelong learning and seamless knowledge transfer for complex, evolving workflows [37]. Advances in model architectures and prompting techniques—such as Chain-of-Thought and SELF-Instruct—show promise for compressing training requirements and operationalizing continual domain adaptation.

8.2 Efficient Tuning and Transfer

The imperative for efficient model adaptation, namely balancing high performance with stringent resource and data constraints, has stimulated research into parameter-efficient tuning, knowledge distillation, and incremental updating strategies. Approaches such as Low-Rank Adaptation (LoRA) and prompt-based fine-tuning significantly reduce the computational and memory footprints of LLMs and RAG systems, enabling effective domain and task transfer at a fraction of the cost compared to full model retraining [36, 37, 55, 60].

Empirical evidence from recommendation and retrieval domains indicates that parameter-efficient tuning not only accelerates model deployment but also facilitates scalable personalization and continual adaptation. When paired with knowledge distillation, these techniques efficiently propagate learned behaviors to lightweight or downstream models [55]. State-of-the-art systems increasingly combine traditional IR pipelines with resource-aware RAG architectures—such as modular index updating and hierarchical retrieval—to reduce redundancy and optimize retrieval quality within data or compute-constrained environments.

The following table summarizes the principal methods and their roles in efficient transfer and adaptation across neural architectures:

In biomedical information extraction settings, multi-task frameworks such as RAMIE integrate instruction fine-tuning with retrieval augmentation to deliver marked resource reductions without sacrificing accuracy, demonstrating the mutual reinforcement of multi-task and retrieval-augmented techniques in minimizing annotation and compute requirements [60]. Domain-specific transfer via continued pretraining and contrastive learning—including enhancements through sparse, dense, or knowledge graph-based retrieval—further allows compact and contextually-grounded adaptation across diverse biomedical and clinical tasks [36, 55].

8.3 Personalization in Retrieval and Recommendation

The field of personalization in retrieval and recommendation has advanced beyond basic pointwise user modeling to encompass sophisticated hierarchical and temporal modeling strategies that effectively capture long-term user preferences and evolving interests. The integration of LLMs into recommender systems, supported by RAG-driven context enrichment and advanced prompt engineering, now enables unprecedented levels of user alignment and explainability across domains [3, 7, 9, 23, 27, 34, 36–39, 42, 46, 50, 53, 55, 58, 60, 64].

Frameworks such as ER2ALM address persistent challenges—such as cold-start scenarios and data sparsity—by fusing LLM capabilities with RAG modules to enrich auxiliary data and denoise user representations. This results in state-of-the-art performance and enhanced resilience in preference mining [3]. Additionally, entity-centric knowledge stores, applied to user interaction histories in search and web platforms, yield lightweight, privacy-preserving user projections that better synchronize LLM outputs with intricate, contextual user preferences [7]. This transition represents a significant move from monolithic user profiling toward modular, user-driven contextualization.

Recent surveys of LLM-based recommendation pipelines highlight several core principles for enhancing user alignment and system trust:

- **Hierarchical preference modeling** enables fine-grained personalization by structuring user histories at multiple temporal or logical scales.
- **Collaborative filtering fusion** leverages shared user behaviors to improve recommendations, even in sparse data regimes.
- **Memory-based prompt scaffolding** incorporates long-term and episodic memory into prompt generation for more contextually aware responses.

Table 5: 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

- **Explainability, fairness, and alignment with domain knowledge** are achieved through the integration of continuous prompt learning, knowledge distillation, and sophisticated regularization.

Ongoing challenges remain, however. Scaling personalization confronts technical obstacles, especially in managing extensive interaction histories, controlling inference latency, and ensuring user privacy—issues that are magnified for ever-larger models with expanded context capacities [39, 53, 55]. Recent results underscore the necessity for parameter-efficient and hybrid adaptation strategies to facilitate real-world deployment. Simultaneously, interpretability, fairness, and ethical safeguards have become critical to ensure alignment with both user goals and broader societal norms.

By synthesizing these developments, it is evident that continual, transfer, and resource-efficient learning, underpinned by advances in modular architectures, parameter-efficient tuning, and nuanced personalization, form the bedrock for next-generation adaptive AI systems. The field's progression hinges on resolving persistent challenges—chiefly catastrophic forgetting, OOD robustness, efficiency under resource constraints, and ethical user alignment—while capitalizing on synergies enabled by recent methodological innovations.

9 Thematic Synthesis 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 catalyzed a transformation in knowledge-intensive AI applications. In particular, RAG models blend large language models (LLMs) with external data repositories—including both structured knowledge graphs and unstructured literature—to mitigate longstanding limitations such as hallucination, outdated content, and lack of traceability inherent to conventional generative systems [3, 13, 22, 23, 30, 33, 37, 40, 47, 48]. This model synergy not only enhances factual correctness, but also enables granular personalization and real-time content updating, effectively serving the demands of specialized domains (e.g., clinical decision support, legal reasoning).

Data augmentation has emerged as a vital mechanism within RAG and allied frameworks. Through approaches such as in-context contrastive learning or pointwise informativeness filtering, models can expand task coverage and robustness—particularly in low-resource and high-variance tasks such as hierarchical classification, intent detection, and few-shot learning [2, 5, 11, 16, 22, 32, 41, 49, 55, 64]. In high-stakes domains like medicine and science, context augmentation coupled with dynamic retrieval proves especially valuable: recent models (e.g., SurgeryLLM, CLEAR) leverage domain-specific guideline integration and entity-centric data chunking to outperform non-augmented baselines in both completeness and efficiency. These trends are substantiated by systematic benchmarks, where meta-analyses demonstrate that RAG-enhanced LLMs consistently achieve significant performance gains (often with odds ratios > 1.35) across a spectrum of biomedical and clinical applications [16].

Contemporary research has also seen the consolidation of explanation and personalization as dominant themes. Techniques such as explicit source citation, stance-based explanations, and contrastive knowledge grounding contribute not only to enhanced user trust, but also to regulatory compliance in high-stakes AI applications [13, 25, 33, 51, 59]. Lightweight personalization strategies—including user-specific knowledge stores and dynamic interaction histories—demonstrably improve contextual relevance, particularly in areas such as information retrieval and question suggestion, without resorting to privacy-invasive profiling [17, 27, 61]. Additionally, contemporary RAG interfaces increasingly incorporate filtering and quality metrics (e.g., factuality scores, stance detection) to proactively mitigate noise and misinformation, which is vital when confronting conflicting or ambiguous evidence streams [25, 26, 30, 38, 45, 50, 52, 60].

9.1.2 Reliability, Explainability, and Security Toward Trustworthy Pipelines. A recurring theme in the literature is the persistent tension between increased model sophistication and operational reliability. While recent pipeline innovations—including debate-based agentic RAGs (such as MADAM-RAG) and multi-stage retrieval combined with re-ranking—have succeeded in reducing hallucination and improving factual completeness, these enhancements often introduce new challenges in orchestration, reproducibility, and explainability [7, 25, 29, 30, 37, 39, 42, 45, 50, 55]. Mechanistic interpretability frameworks have become essential, providing fine-grained diagnostic capabilities that allow practitioners to trace causal pathways and intervene directly in parametric neural information retrieval (IR) systems. Such capacities are especially critical as these systems frequently underpin pivotal decision-support processes in healthcare and law [24, 34, 36].

Security and adversarial robustness remain significant open challenges. Published studies highlight measurable vulnerabilities of dense and neural ranking models to out-of-distribution (OOD) data and coordinated adversarial attacks; hence, trustworthiness, transparency, and continuous adaptation have emerged as foundational requirements for deployments in mission-critical contexts [13, 20, 28, 29, 37, 55, 62, 63]. Best practices for trustworthy deployment now typically include:

- Ongoing monitoring of retrieval quality,
- Rigorous quality control (including robust filtering and re-ranking steps),

- Privacy-preserving personalization (notably via aggregate projection-based knowledge stores rather than fully individualized user profiles).

[17, 27, 46, 50]

Explainability is increasingly being operationalized at the system interface layer, through mechanisms such as traceable source grounding, contrastive explanations tailored to explicit user goals, and hybrid architectures that combine computational argumentation with knowledge graphs. The convergence of transformer-based retrieval with computational argumentation and graph-based reasoning methods points to the emergence of more user-centric, explainable, and multimodal AI systems [3, 5, 7, 8, 28, 33, 51, 53, 59].

9.1.3 Cross-Modal, Unified Learning and Workflow Innovation. A central and intensifying trend is the generalization of RAG, context-augmented, and contrastive approaches beyond language, paving the way for unified methodologies encompassing vision, multimodal content, and graph-structured data [3–5, 19, 33, 37, 46–48, 50]. Recent developments in cross-modal retrieval and hashing frameworks exploit synergies between diverse modalities—addressing the specific heterogeneities that arise, for example, in aligning subjective textual content with objective visual imagery (as exemplified by GCDH and multimodal transfer architectures) [2, 19]. Noteworthy innovations include retrieval-pretrained transformers (RPT) and unified pretraining regimes, which jointly optimize retrieval and generation for long-range semantic comprehension. These yield measurable improvements in model perplexity and retrieval precision on complex scientific and legal corpora [5, 21, 33, 37].

Workflow optimization, another area of active research, is facilitated by contextual integration of external tools and map-reduce-inspired strategies—partitioning context and leveraging tool APIs for tasks such as experimental design or clinical procedure planning, as implemented in CALMS and BriefContext [17, 39, 53]. Such integrations not only decrease hallucination rates and improve operational completeness, but also expedite domain-specific knowledge transfer. This development marks a fundamental shift from passive knowledge extraction to proactive, tool-augmented reasoning [17, 30]. Complementing these advances, harmonized evaluation protocols—such as the S.C.O.R.E. framework and GUIDE-RAG staging—are advancing performance standardization and facilitating inter-study comparability [16, 30, 50].

Table 6 provides a concise, modality-centric overview of representative innovations and their primary domains of application.

9.2 Future Directions

9.2.1 Toward Unified, Multimodal, and Cross-Domain Frameworks. The ongoing trajectory for knowledge-augmented language models is pointedly aligned with the emergence of unified frameworks, capable of seamless cross-modal and cross-domain information integration [13, 22, 37, 40]. These architectures are expected to harmonize disparate data sources—including text, images, graphs, and personalized user histories—enabling universal retrieval and generative reasoning. Recent progress is seen in experimental systems that, for example:

- Connect graph-based and textual knowledge for dialogue agents,
- Extract and encode multimodal semantics for robust retrieval,
- Aggregate heterogeneous, domain-specific corpora (e.g., medical images, chemical graphs, user activity logs) for comprehensive AI-driven assistance.

[3, 5, 27, 33, 39, 48, 51]

Achieving dynamic, multilingual, and multimodal stream processing while maintaining explainability and efficiency will demand further breakthroughs in representation learning, domain-specific adaptation, and interpretability toolkits. Integration of distributed knowledge spaces with RAG pipelines holds the promise of delivering trustworthy, interoperable, and secure access to high-quality data, an imperative in both open-access and regulated domains [13].

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 OOD scenarios,
- End-to-end deployment efficacy, including latency, scalability, and regulatory compliance.

The systematic development and adoption of such resources is critical for empirical validation and progress toward reliable, real-world systems [16, 55, 60].

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

- **Scalability:** Computation and knowledge management at scale continues to limit practical deployment of end-to-end, joint retrieval-generation models across heterogeneous, large-scale data ecosystems [5, 13, 19, 20, 23, 25, 32, 33, 37, 38, 43, 50, 62, 63].
- **Data Scarcity:** The lack of high-quality annotation and curated datasets especially hampers progress in specialized or rare domains. While LLM-assisted data synthesis provides some respite, it cannot fully replace real, expert-annotated corpora [3, 11, 16, 21, 26, 30, 34, 41, 64].
- **Robustness:** Adversarial risks—stemming from misinformation, conflicting evidence, or environmental noise—are unsolved, and demand ongoing innovation in both algorithmic robustness and evaluation paradigms [7, 20, 28, 29, 33, 42, 62, 63].
- **Ethics, Privacy, and Compliance:** These remain largely unaddressed frontiers. While efforts in privacy-by-design,

Table 6: Representative Innovations in Knowledge-Augmented AI: Modalities and Applications

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

fairness-aware algorithmic prompting, and transparent citation are emerging, robust, standardized frameworks are still lacking in regulated domains such as healthcare, science, and law, wherein AI-expert outputs directly impact critical decisions and human welfare [9, 13, 22, 27, 28, 35, 46, 55, 61].

In summary, the field stands at a pivotal juncture: advances in RAG, context augmentation, and contrastive architectures have established new benchmarks for reliability, explainability, and performance in knowledge-intensive AI. Yet, meaningful, scalable deployment in real-world, high-stakes settings necessitates the development of integrative solutions—including unified multimodal frameworks, empirically robust evaluation resources, and comprehensive approaches to ethical, technical, and regulatory challenges.

10 Conclusion and Strategic Outlook

10.1 Synthesis Across Methods and 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, 39]. These efforts are evident in the modernization of retrieval pipelines, which increasingly incorporate user-centric personalization—leveraging interaction histories, lightweight knowledge graphs, and dynamic embeddings—to achieve greater contextual relevance across both general web search and specialized clinical settings [10, 53, 63].

Context augmentation—encompassing retrieval-augmented generation (RAG) frameworks, knowledge graph-driven models, and user history integration—is vital for mitigating LLM hallucinations and overcoming the limitations of closed-book systems [32, 43]. By infusing model prompts with retrieved, verifiable knowledge, both scientific and clinical applications benefit from improvements in accuracy and interpretability [58, 62]. The healthcare sector exemplifies this trend: integrating codified guidelines, structured health records, and multimodal clinical data enables LLMs to deliver outputs that are both consistent and safe, exceeding what static, non-augmented models can offer [35, 48, 59]. This methodological rigor yields tangible improvements in patient safety and cultivates clinician trust, with retrieval-augmented frameworks such as SurgeryLLM and CLEAR demonstrating superior diagnostic accuracy, documentation quality, and alignment with established standards of care [35, 42, 43].

Parallel advances in contrastive learning and data augmentation have been transformative for recommendation systems and intent detection. Multi-level contrastive learning methods aggregate item-wise, batch-wise, and sequence-wise signals, thereby improving data efficiency and cold-start resilience in sequential recommender systems [54, 56]. Synthetic data generation with open-source LLMs (e.g., LLaMA, Alpaca) has proven especially beneficial in privacy-sensitive and label-scarce environments, expanding the diversity and robustness of training data while maintaining user confidentiality [14]. Additionally, developments in multimodal integration—spanning cross-modal retrieval and hybrid graph/neural models—have bolstered representation learning across text, image, and structured domains. This progress drives high-impact applications such as industrial defect detection and biomedical literature navigation [31, 47].

Personalization strategies now emphasize lightweight, privacy-preserving models that enrich LLMs with user-specific knowledge repositories, aggregate behavioral profiles, and context-derived features to maximize output relevance and utility [45, 63]. This trend is acutely significant in domains where compliance, trust, and user agency are crucial, including recommendation, healthcare, and legal AI. Furthermore, cross-domain and multimodal integration—achieved through transfer learning and graph-augmented architectures—expands the scope and robustness of retrieval-augmented models, particularly where data is sparse, noisy, or distributed across heterogeneous infrastructures [47, 48, 50].

Despite these successes, several core challenges persist:

- **Retrieval bottlenecks** in complex, highly related corpora remain consequential [9].
- Model sensitivity to **context length** and **data density** creates vulnerabilities [52].
- Limitations exist in **data augmentation** regarding nuanced or context-heavy tasks [18].
- Scaling RAG frameworks to emerging modalities and dynamic regulatory requirements is difficult [3, 8].
- Synthetic and augmented data are helpful but insufficient for achieving contextual depth, necessitating ongoing qualitative review [18].

Strategically, the community must prioritize enhanced evaluation methods, responsible and user-centric research, and broad interdisciplinary collaboration. For evaluation, emphasis should be on metrics capturing OOD generalization, multi-agent debate, and data diversity, moving beyond insular benchmarks to simulate

genuine deployment pressures and user heterogeneity [7, 23, 39]. Responsible research requires transparency in retrieval provenance, automated audit trails, user-driven customization, and adherence to formal compliance frameworks addressing privacy and explainability [3, 17, 34, 38]. Moreover, interdisciplinary engagement involving informatics, regulatory science, ethics, and human-computer interaction will be key in translating methodological innovations into scalable, trustworthy automation—particularly within healthcare, legal, and public sector environments [10, 29, 33, 45].

Best practices recommend the following:

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

These practices are prerequisites for the responsible adoption of AI in high-stakes settings, ensuring that the balance between scalable automation and human oversight is continually recalibrated to protect both model utility and user trust.

10.2 Vision for Real-World 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 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].

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].

Ultimately, 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 this outcome requires sustained interdisciplinary synthesis and unwavering dedication to ethics, transparency, and scientific rigor across all methods and domains.

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.

[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