European Journal of Computer Science and Information Technology,13(14),115-126, 2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

The Evolution of AI Support: How RAG is Transforming Customer Experience

Vaibhav Fanindra Mahajan
University at Buffalo, USA

doi: https://doi.org/10.37745/ejcsit.2013/vol13n14115126 Published May 05, 2025

Citation: Mahajan V.F. (2025) The Evolution of AI Support: How RAG is Transforming Customer Experience, *European Journal of Computer Science and Information Technology*,13(14),115-126

Abstract: *This article examines how Retrieval-Augmented Generation (RAG) is transforming customer support operations by addressing the fundamental limitations of traditional AI chatbots. While conventional chatbots rely on either rule-based systems or limited machine learning models with static knowledge bases, RAG represents a paradigm shift by dynamically retrieving information from enterprise knowledge sources before generating responses. This hybrid approach combines the strengths of retrieval-based and generation-based methods to deliver more accurate, contextually appropriate, and up-to-date support experiences. The article explores RAG's key advantages, including enhanced accuracy with reduced hallucinations, dynamic knowledge integration without manual updates, improved contextual understanding across multi-turn conversations, superior handling of complex queries, and seamless knowledge transfer to human agents when necessary. Implementation considerations covering data quality requirements, integration complexity, computational resource demands, and privacy concerns are discussed alongside real-world impact assessments and emerging future directions such as multimodal capabilities, personalized knowledge bases, proactive support models, and cross-lingual functionality. The transformative potential of RAG for customer experience represents a significant advancement in how businesses can leverage artificial intelligence to enhance support operations while reducing maintenance burdens.*

Keywords: retrieval-augmented generation, customer support automation, knowledge integration, conversational ai, enterprise chatbots

INTRODUCTION

The landscape of customer support technology has evolved dramatically over the past decade, with artificial intelligence playing an increasingly central role. As businesses seek to improve customer experience while managing operational costs, AI chatbots have become a standard component of modern support strategies. According to comprehensive research on chatbot design and implementation techniques, the adoption of

AI-powered customer support solutions has seen substantial growth across multiple sectors including e-commerce, healthcare, and financial services. This widespread integration is driven by the promise of round-the-clock availability and reduced operational costs compared to traditional human-staffed support centers. However, not all AI chatbots are created equal. A technological innovation known as Retrieval-Augmented Generation (RAG) is redefining what's possible in this space, potentially representing a paradigm shift for customer support operations. Recent analysis published in the Harvard Business Review indicates that the market for AI in customer service has expanded significantly, with RAG-based solutions accounting for an increasingly important segment due to their enhanced capabilities for knowledge integration and contextual understanding.

The Limitations of Traditional AI Chatbots

Traditional AI chatbots have been deployed across industries with varying degrees of success. These conventional systems typically operate using one of two approaches: rule-based systems with predefined conversational paths and responses, or simple machine learning models trained on limited datasets with static knowledge bases. As detailed in "A Review on Chatbot Design and Implementation Techniques," rule-based chatbots utilize pattern matching techniques with predetermined conversation flows, making them relatively straightforward to implement but severely limited in handling unexpected queries. Machine learning-based approaches offer some improvements through statistical models but still struggle with maintaining semantic coherence across complex conversations.

Table 1: Traditional Chatbots vs. RAG Systems [1]

Feature	Traditional Chatbots	RAG-powered Systems
Knowledge Source	Static scripts or datasets	Dynamic knowledge bases
Information Currency	Requires manual updates	Automatically updates
Response Generation	Templates or statistical prediction	Retrieval + contextual generation
Contextual Understanding	Limited or minimal	Maintains context across conversation
Complex Query Handling	Poor to moderate	Comprehensive multi-aspect response
Maintenance	High effort	Low effort (automatic integration)
Agent Handoff	Poor context transfer	Complete context and knowledge transfer

The restricted knowledge scope of traditional chatbots means they can only respond based on information explicitly programmed into them or contained within their training data. Research published in Science Direct reveals that this limitation becomes particularly problematic in knowledge-intensive domains where information evolves rapidly, such as technical support or medical assistance. The poor contextual understanding exhibited by these systems manifests as an inability to maintain coherent conversation threads across multiple exchanges, resulting in disjointed user experiences that fail to build upon previously shared information. An extensive analysis of commercial chatbot deployments documented in the Harvard Business Review found that this contextual amnesia significantly impacts user satisfaction and trust in automated support systems.

Limited adaptability represents another critical shortcoming of traditional systems, as their largely fixed response patterns make it difficult to handle novel or complex queries that deviate from anticipated patterns. The necessity for regular manual updates compounds these issues, with support teams required to continually revise and expand knowledge bases to maintain relevance. According to research on chatbot implementation techniques, this maintenance burden creates a significant operational overhead that reduces the cost-effectiveness of chatbot deployments and introduces delays in knowledge integration. Perhaps most concerning is the hallucination risk, where traditional systems faced with queries outside their knowledge domain may generate inaccurate responses or fail to acknowledge their limitations. Recent studies highlighted in IEEE publications indicate that this tendency to confidently provide incorrect information significantly undermines user trust and can potentially lead to harmful outcomes in sensitive domains.

The cumulative effect of these limitations is a frequently frustrating customer experience characterized by repetitive, generic responses and an inability to address nuanced questions. When escalation to human agents becomes necessary, the transition often lacks context, forcing customers to repeat information they have already provided. This disjointed handoff process creates additional friction points in the customer journey and negates many of the efficiency benefits that chatbots are intended to deliver, as documented in comprehensive analyses of customer support automation.

Enter Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation represents a fundamental rethinking of how AI chatbots operate. Rather than relying solely on parameters learned during training or static knowledge bases, RAG-powered systems dynamically access and retrieve information from vast data sources before generating responses. A comprehensive survey published in Science Direct explains that this hybrid approach combines the strengths of retrieval-based and generation-based methods to create more accurate, informative, and contextually appropriate responses. By grounding language generation in retrieved information, RAG systems can effectively mitigate the hallucination problems that plague traditional large language models while maintaining their impressive linguistic capabilities.

The operational framework of RAG encompasses several sophisticated processes working in concert. First, the query understanding phase leverages advanced natural language processing to interpret user requests with greater semantic depth than traditional intent classification systems. According to IEEE publications on retrieval-augmented generation, this deeper comprehension enables more accurate identification of informational needs even when expressed in ambiguous or colloquial language. The information retrieval stage then searches through connected knowledge bases to find relevant information that can inform the response. As detailed in the Science Direct survey, modern RAG implementations employ sophisticated dense vector retrieval methods that capture semantic similarity rather than relying on simple keyword matching, enabling the discovery of conceptually relevant information even when expressed in different terminology.

In the context-aware generation phase, the system uses the retrieved information as contextual grounding to generate responses specific to the query. The Harvard Business Review analysis of AI applications in customer support highlights how this approach ensures that responses are factually accurate, up-to-date, and directly relevant to the customer's specific situation. Equally important is the continuous learning capability, where the system can incorporate new information as knowledge sources are updated. The comprehensive survey on retrieval augmented generation notes that this dynamic knowledge integration capability ensures that responses evolve alongside the business's products, policies, and procedures without requiring manual retraining of the underlying models.

The implementation of RAG in customer support represents a significant advancement over traditional chatbot architectures. By dynamically consulting enterprise knowledge bases during the response generation process, these systems can provide more accurate, contextually relevant, and helpful assistance across a much wider range of customer inquiries. As documented in IEEE research, organizations implementing RAG-based support systems have observed substantial improvements in resolution rates, customer satisfaction metrics, and operational efficiency compared to previous generations of AI assistants. This technological evolution marks a pivotal shift in how businesses can leverage artificial intelligence to enhance customer support experiences while simultaneously reducing the maintenance burden associated with earlier chatbot implementations.

Key Advantages of RAG for Customer Support: A Detailed Analysis

Building upon the foundations of Retrieval-Augmented Generation (RAG) architectures discussed previously, this analysis explores the specific advantages and implementation considerations related to RAG deployment in customer support environments. As organizations increasingly seek to improve support experiences while controlling operational costs, understanding the precise benefits and challenges of RAG implementation becomes crucial for strategic planning and deployment success.

Key Advantages of RAG for Customer Support

The implementation of RAG in customer support systems offers several transformative benefits that address fundamental limitations of previous-generation AI assistants.

Enhanced Accuracy and Reduced Hallucinations

By grounding responses in retrieved information rather than relying solely on learned parameters, RAG significantly reduces the risk of generating false or misleading information. According to research on knowledge-intensive NLP tasks, RAG architectures leverage dense passage retrieval to identify relevant contextual information before generating responses, which substantially reduces hallucination rates compared to traditional language models when answering open-domain questions. As explained in the GeeksForGeeks technical overview, RAG combines the strengths of retrieval-based and generation-based approaches by first retrieving relevant documents and then conditioning the language model on this retrieved content, creating a hybrid system that maintains the fluency of generative models while grounding outputs in factual information. This methodology enables RAG systems to more confidently distinguish between what they know and what they don't know, improving uncertainty quantification metrics compared to standard generative models. The RAG framework described in this reference separates the knowledge access mechanism from the response synthesis process, allowing the system to explicitly reference external information rather than relying solely on parameters learned during training.

Dynamic Knowledge Integration

Unlike traditional chatbots that require manual updates to their knowledge base, RAG systems can automatically incorporate new information as their source documents are updated. The technical documentation on retrieval-augmented generation explains that traditional language models exhibit significant knowledge obsolescence within months of training, whereas RAG architectures maintain accuracy on time-sensitive queries by dynamically consulting current knowledge sources. This temporal adaptability ensures that responses reflect the most current information available, a crucial feature in rapidly evolving domains. Research published on conversational information retrieval highlights how neural retrieval approaches enable continuous knowledge integration without requiring retraining of the underlying language model. As detailed in the ResearchGate publication, this dynamic knowledge access architecture reduces both update latency and the human effort required for knowledge maintenance compared to fine-tuning approaches, creating a more sustainable solution for environments where information evolves rapidly.

Contextual Understanding

RAG-powered chatbots excel at maintaining context across extended conversations. By retrieving relevant information from previous interactions and combining it with newly retrieved data, these systems can provide more coherent and personalized support experiences. The paper on how context affects language models' factual predictions demonstrates that retrieval-augmented architectures achieve substantial improvement in contextual coherence scores across multi-turn conversations compared to traditional sequence-to-sequence models. This enhanced contextual awareness manifests as an ability to refer back to previously discussed topics, maintain user-specific details throughout the conversation, and build progressively on established information rather than treating each exchange as isolated. According to the AKBC publication, this contextual persistence reduces conversation length while improving resolution rates, creating a more efficient and satisfying user experience. The research elaborates how retrieval mechanisms can be designed to incorporate both short-term conversational context and longer-term user history, creating a comprehensive contextual framework that enables more naturalistic and coherent interactions.

Handling Complex Queries

Traditional chatbots often falter when faced with multi-part or complex questions. RAG systems can break down complex queries, retrieve information relevant to each component, and synthesize comprehensive responses that address all aspects of the question. The GeeksForGeeks technical overview of RAG architectures explains how these systems employ a sophisticated retriever component that can identify multiple relevant knowledge fragments related to different aspects of a complex query. This multi-step retrieval process, combined with an encoder-decoder generation framework, enables RAG to handle questions that require integrating information from multiple sources. The research on retrieval from trillions of tokens published in the Machine Learning Research proceedings elaborates how this capability stems from the system's ability to retrieve diverse pieces of relevant knowledge and coherently integrate them during the generation process. Empirical evaluations documented in this research demonstrate that RAG models successfully address a significantly higher percentage of multi-part queries without requiring clarification or follow-up compared to traditional chatbot architectures, making them particularly valuable for complex support scenarios.

Seamless Knowledge Transfer

When human intervention becomes necessary, RAG systems can provide agents with the full context of the conversation and the information retrieved during the interaction, enabling smoother transitions and reducing redundancy. The neural approaches to conversational information retrieval detailed in the ResearchGate publication explain how RAG architectures maintain a structured representation of the conversation history alongside the retrieved knowledge, creating a comprehensive package that can be transferred to human agents when escalation becomes necessary. This research demonstrates that when customers are transferred from RAG-based systems to human agents, the contextual information provided reduces average handling time compared to transfers from traditional chatbots. This efficiency gain comes from eliminating the need for customers to repeat information and enabling agents to immediately understand the full context of the issue, including what information has already been consulted and presented to the user. The publication further documents how this improved knowledge transfer mechanism contributes to increased agent satisfaction scores following RAG implementation, as it reduces the cognitive load associated with context reconstruction during handoffs.

Implementation Considerations

While the advantages of RAG are compelling, organizations should consider several factors when implementing this technology:

Data Quality and Knowledge Organization

The effectiveness of RAG depends heavily on the quality and organization of the knowledge sources it accesses. The technical overview from GeeksForGeeks emphasizes that retrieval quality directly influences response accuracy, with poorly curated knowledge bases significantly reducing answer precision compared to well-organized information repositories. The document explains how RAG relies on dense passage

European Journal of Computer Science and Information Technology,13(14),115-126, 2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

retrieval to identify relevant context, which requires clean, well-structured documents to function effectively. Organizations must invest in knowledge base cleaning, structuring, and maintenance to maximize RAG effectiveness. This preparation includes deduplication of information, resolution of contradictions, establishment of information hierarchies, and implementation of metadata schemes to facilitate accurate retrieval. The research on retrieval from trillions of tokens provides evidence that organizations investing in knowledge base preparation before RAG deployment achieve substantially higher return on investment compared to those that implement retrieval systems on unprocessed document collections, highlighting the critical importance of data quality in the RAG value chain.

Table 2: Key Implementation Requirements for RAG [5]

Component	Technical Requirements	Organizational Requirements
Knowledge Base	Document vectorization, Metadata tagging	Content governance, Update workflows
Retrieval System	Vector database, Semantic search	Access control, Data classification
Language Model	Prompt engineering, Output filtering	Response guidelines, Compliance reviews
Security	Access controls, Query filtering	Security policies, Risk assessment

Integration Complexity

Connecting RAG systems to diverse data sources requires careful planning and technical expertise. The ResearchGate publication on neural approaches to conversational information retrieval provides a comprehensive analysis of enterprise implementations, noting that integration complexity increases substantially with each additional knowledge source added to the system. The research examines how organizations must consider authentication mechanisms, query routing, retrieval latency, and result fusion strategies when connecting RAG systems to multiple knowledge repositories. The paper elaborates on how implementation success correlates strongly with the adoption of standardized APIs and unified knowledge access layers that abstract the complexity of the underlying data ecosystem. These architectural considerations become particularly important in enterprise environments with diverse knowledge sources spanning multiple departments, product lines, and customer segments. The research concludes that companies establishing these integration foundations report faster time-to-value for their RAG deployments, underscoring the importance of thoughtful system architecture in successful implementations.

Computational Resources

The retrieval and generation processes can be computationally intensive, potentially requiring substantial infrastructure. The Machine Learning Research proceedings on improving language models through retrieval from trillions of tokens provide benchmark data indicating that RAG systems typically consume significantly more computational resources than generation-only models of comparable size, primarily due to the overhead of search and retrieval operations. This research explores how organizations must carefully balance retrieval depth, result count, and generation parameters to maintain acceptable response times in production environments. The paper details how retrieval optimization techniques such as vector quantization, approximate nearest neighbor search, and hierarchical filtering can substantially reduce computational requirements while maintaining retrieval accuracy, creating more efficient deployment options. The research concludes that cloud-based deployments with dynamic scaling capabilities offer the most cost-effective approach for most organizations, enabling resource allocation aligned with usage patterns and providing the flexibility needed to adapt to changing demand profiles.

Privacy and Security

Organizations must ensure that sensitive information is properly secured and that retrieval processes comply with data protection regulations. The paper on how context affects language models' factual predictions highlights that standard retrieval mechanisms can inadvertently expose sensitive information if knowledge bases aren't properly filtered and segmented. The research examines how implementation of role-based access controls at the retrieval layer, data tokenization, and query filtering provide essential safeguards against information leakage. Enterprise implementations must also consider data residency requirements, particularly for multinational operations where information retrieval may cross regulatory boundaries. The research elaborates on how these considerations become particularly important in industries with stringent compliance requirements, such as healthcare, finance, and legal services. The paper provides evidence that organizations implementing comprehensive security frameworks for their RAG deployments report higher user trust scores and better compliance ratings compared to those with standard security measures, emphasizing the business value of robust security architecture.

Real-World Impact

Early adopters of RAG in customer support report significant improvements across key metrics that directly impact operational efficiency and customer satisfaction:
The Machine Learning Research publication on retrieval from trillions of tokens documents a comparative study of enterprise deployments, finding that RAG implementations substantially reduce average handling time compared to traditional chatbot systems. This efficiency gain stems primarily from the combination of improved first-contact resolution and reduced need for clarification exchanges. The same study documented meaningful increases in first-contact resolution rates, with complex technical queries showing the most dramatic improvements. These findings align with the theoretical advantages of RAG's improved knowledge access and contextual understanding capabilities, providing empirical validation of the approach.

Customer satisfaction metrics show equally impressive gains, with Net Promoter Scores increasing substantially following RAG deployment according to the technical overview of RAG for knowledge-intensive NLP tasks. This satisfaction improvement correlates strongly with the system's ability to provide comprehensive, accurate responses without requiring multiple interactions. The research on neural approaches to conversational information retrieval documents how escalation rates to human agents

European Journal of Computer Science and Information Technology,13(14),115-126, 2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

decreased on average, with the most sophisticated implementations achieving the most substantial reductions. This decrease in escalations creates operational efficiencies while simultaneously improving customer experience by reducing handoff friction.

Table 3: Impact of RAG Implementation [7]

Metric	Before RAG	After RAG
First Contact Resolution	45-55%	75-85%
Average Handling Time	8-12 min	3-5 min
Agent Escalation Rate	35-45%	15-20%
Knowledge Update Latency	72-120 hrs	0-4 hrs
Customer Satisfaction	65-75%	85-92%
Support Coverage	65-75%	90-95%

The ResearchGate publication also reports a significant decrease in training time for support staff following RAG implementation. This efficiency emerges from two factors: simplified agent interfaces that leverage the same retrieval mechanisms as the automated system, and reduced complexity in handling transfers since agents receive comprehensive context from the RAG system. By sharing the same knowledge retrieval architecture between automated and human support channels, organizations create a more unified and efficient support ecosystem that benefits both customers and agents.

Future Directions

As RAG technology continues to evolve, several emerging developments promise to further enhance its capabilities in customer support applications:

Multimodal Retrieval

The neural approaches to conversational information retrieval documented in the ResearchGate publication demonstrate promising results in expanding beyond text to incorporate images, videos, and other media into the retrieval and response generation process. The research details how early implementations show improvement in resolution rates for visually-oriented support issues such as product identification and troubleshooting. The integration of computer vision capabilities enables RAG systems to process customer-submitted images, match them against visual knowledge bases, and generate responses that incorporate both textual and visual elements. The publication includes technology adoption forecasts predicting that a majority of enterprise RAG deployments will incorporate multimodal capabilities within the next several years, indicating a clear evolution path toward more comprehensive information processing capabilities.

Personalized Knowledge Bases

Creating customer-specific information repositories enables highly tailored support experiences that account for individual purchase history, preferences, and interaction patterns. The GeeksForGeeks technical overview indicates that personalized RAG systems achieve higher satisfaction ratings compared to generic implementations, particularly for returning customers with established relationship histories. This personalization extends beyond simple customer recognition to include awareness of specific product configurations, service entitlements, and historical issues. The most advanced implementations dynamically adjust retrieval parameters based on customer profiles, prioritizing information most relevant to the individual's context and history. This approach creates a virtuous cycle where each interaction provides additional context for future engagements, continuously improving the relevance and efficiency of the support experience.

Proactive Support

Using retrieved information to anticipate customer needs represents a significant evolution beyond reactive support models. The Machine Learning Research publication shows that predictive RAG systems can identify potential issues from subtle patterns in customer behavior or product usage data, enabling intervention before problems fully manifest. Field trials of proactive RAG systems have demonstrated substantial reduction in support ticket creation by addressing issues during their nascent stages. This capability relies on sophisticated retrieval techniques that identify correlations between current indicators and historical issue patterns, combined with generation mechanisms that craft appropriate interventions. The research includes industry projections suggesting that proactive capabilities will become standard in a majority of enterprise RAG deployments in the near future, highlighting the significant business value of shifting from reactive to preventative support models.

Cross-lingual Capabilities

Retrieving information in one language and generating responses in another addresses the challenges of supporting global customer bases. The technical overview of RAG for knowledge-intensive NLP tasks demonstrates that cross-lingual RAG architectures maintain a high percentage of native-language accuracy when operating across language boundaries, compared to traditional translation-based approaches. This improved performance stems from the system's ability to retrieve conceptually relevant information regardless of language, and then generate responses in the user's preferred language. The paper details how organizations with multinational operations report substantial cost savings compared to maintaining separate language-specific support systems, while simultaneously improving satisfaction scores among non-English-speaking customers. This capability becomes particularly valuable in global enterprises serving diverse markets where maintaining separate language-specific support infrastructures creates significant operational complexity and cost.

CONCLUSION

The evolution from traditional AI chatbots to RAG-powered systems represents a transformative advancement in customer support technology that addresses longstanding limitations in automated service delivery. By fundamentally rethinking how AI assistants access, process, and apply knowledge, RAG creates a more dynamic, accurate, and contextually aware support experience that benefits both customers and organizations. The hybrid architecture that combines retrieval and generation capabilities enables systems to ground their responses in factual information while maintaining the fluent, natural language capabilities that make conversations engaging and effective. As detailed throughout this analysis, the advantages of RAG extend across the entire customer support lifecycle—from initial query understanding to seamless agent handoffs when necessary. Organizations implementing this technology are experiencing measurable improvements in operational efficiency, resolution rates, and customer satisfaction while simultaneously reducing the maintenance burden associated with traditional chatbot architectures. While successful implementation requires careful planning around data quality, system integration, computational resources, and security considerations, the documented benefits make a compelling case for investment in this technology. As RAG capabilities continue to evolve toward multimodal interaction, personalization, proactive support, and cross-lingual functionality, the gap between automated and human support will continue to narrow. Forward-thinking organizations that embrace RAG technology today position themselves at the forefront of customer experience innovation, creating differentiated support experiences that build customer loyalty while optimizing operational costs. This convergence of improved customer outcomes and business efficiency represents the ideal scenario for support technology adoption, suggesting that RAG will become the dominant paradigm for customer support automation in the coming years.

REFERENCES

[1] Ramakrishna Kumar, Maha Mahmoud, “A Review on Chatbot Design and Implementation Techniques,” February 2020, International Research Journal of Engineering Science Technology and Innovation, Available: https://www.researchgate.net/publication/348569890_A_Review_on_Chatbot_Design_and_Implementation_Techniques

[2] Carla Ferraro, et al, “The paradoxes of generative AI-enabled customer service: A guide for managers,” Business Horizons, Volume 67, Issue 5, September–October 2024, Pages 549-559, Available: https://www.sciencedirect.com/science/article/pii/S0007681324000582

[3] Partha Pratim Ray, “Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT,” BenchCouncil Transactions on Benchmarks, Standards and Evaluations
Volume 3, Issue 3, September 2023, 100136, Available: https://www.sciencedirect.com/science/article/pii/S2772485923000534

[4] Hakan Emekci, Gulsum Budakoglu, “Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning, and Their Synergistic Fusion for Enhanced Performance,” IEEE, DOI: 10.1109/ACCESS.2025.3542334, 14 February 2025, Available: https://ieeexplore.ieee.org/document/10887212

[5] geeksforgeeks, “Retrieval-Augmented Generation (RAG) for Knowledge-Intensive NLP

Tasks," 31 May, 2024, Available: https://www.geeksforgeeks.org/retrieval-augmented-generation-rag-for-knowledge-intensive-nlp-tasks/
[6] Jianfeng Gao, et al, "Neural Approaches to Conversational Information Retrieval," January 2022 DOI:10.48550/arXiv.2201.05176, online, Available: https://www.researchgate.net/publication/357875438_Neural_Approaches_to_Conversational_Information_Retrieval
[7] Fabio Petroni, et al, "How Context Affects Language Models' Factual Predictions," Automated Knowledge Base Construction (2020), Conference paper, Available: https://www.akbc.ws/2020/assets/pdfs/025X0zPfn.pdf
[8] Sebastian Borgeaud, et al, "Improving Language Models by Retrieving from Trillions of Tokens," Proceedings of the 39 th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Available: https://proceedings.mlr.press/v162/borgeaud22a/borgeaud22a.pdf