检索增强生成 (RAG) 的最新进展

Owner	志 志刚 王
Tags	

引言：检索增强生成的崛起

检索增强生成 (RAG) 作为一项关键技术，已在人工智能领域占据核心地位，通过整合外部知识来增强大型语言模型 (LLM) 的准确性和可靠性.1 这项技术有效地弥补了 LLM 在信息更新、领域知识覆盖以及潜在的“幻觉”问题等方面的不足.3 RAG 的一个显著优势在于其能够提供生成内容的出处，如同研究论文中的脚注一样，使用户可以核实信息的真实性，从而建立信任.6 此外，相较于对 LLM 进行再训练或大规模微调，RAG 提供了一种更具成本效益的知识集成方案.3 本报告旨在全面概述 RAG 的最新进展，涵盖其定义、架构演进、知识库优化、评估方法、行业应用、新兴工具与框架以及未来趋势。

分析表明，RAG 的核心价值在于其能够利用外部数据来提升 LLM 的性能，尤其是在需要最新或特定领域知识的场景中。通过检索相关信息并将其融入生成过程中，RAG 不仅提高了答案的准确性，也增强了用户对生成内容的信任。此外，RAG 的成本效益使其成为各种规模组织采用先进人工智能技术的实用选择。

理解 RAG 的基本原理

一个典型的 RAG 系统由检索模型、知识库和生成模型三个核心组件构成.2 首先，需要从知识库中提取与用户查询相关的外部信息。这个过程通常涉及将知识库中的文档和用户查询都转换为向量嵌入.2 向量嵌入是一种数值表示形式，能够捕捉文本的语义含义。借助向量数据库，系统可以高效地进行相似性搜索，从而快速检索到与用户查询在语义上最相关的信息.2 随后，系统会将检索到的相关上下文信息添加到 LLM 的提示中，以指导其生成最终的答案.2 在检索阶段，有多种技术可供选择，包括语义搜索、关键词搜索以及结合两者优势的混合搜索.2

分析表明，RAG 系统的模块化设计是其能够不断进步的关键。通过将检索和生成过程分离，研究人员和开发人员可以专注于改进每个组件的性能。此外，混合搜索技术的应用体现了在信息检索中结合不同方法的优势，以应对各种用户查询和数据场景。

RAG 系统的架构创新与增强

近年来，RAG 系统的架构出现了显著的创新和增强，旨在进一步提升其性能和适用性。

- **GraphRAG:** 这是一种将图结构数据集成到检索增强生成中的新型扩展，旨在通过增强知识检索和生成来改进推理能力.15 GraphRAG 包含查询处理、检索、组织和生成等关键组件。
- **VideoRAG:** 与主要依赖文本或将视频转换为文本描述的现有方法不同，VideoRAG 能够动态检索相关视频，并利用视频中的视觉和文本信息.15 它利用大型视频语言模型 (LVLM) 直接处理视频内容，从而改进检索和生成效果。
- **Agentic RAG:** 通过集成自主 AI 代理来增强传统 RAG，从而改进检索策略、上下文优化和工作流程适应性.15 Agentic RAG 利用了诸如反思、规划和工具使用等代理设计模式。
- **链式检索增强生成 (CoRAG):** 迭代地检索和推理信息，以解决多跳问题.15 CoRAG 通过中间检索链进行查询重构。
- **多模态 RAG:** 将 RAG 的处理能力扩展到文本以外的各种数据格式，例如图像、音频和视频.15
- **自适应检索:** 根据用户意图和查询复杂性动态调整检索策略.17
- **推测性 RAG:** 使用一个较大的通用 LM 来有效地验证由一个较小的专业 LM 并行生成的多个 RAG 草稿.19

分析表明，RAG 架构的演进趋势是朝着更专业化和智能化的方向发展。针对特定数据类型和复杂任务定制的架构，以及能够动态调整检索策略和利用多种模型协同工作的方法，预示着 RAG 技术在未来将能够处理更广泛和更复杂的应用场景。

构建和维护有效的 RAG 知识库

对于任何 RAG 系统而言，构建和维护一个有效的知识库至关重要。知识库的质量直接影响着 RAG 系统检索到的信息的准确性和相关性.20 这就要求知识库中的数据必须是高质量、一致且最新的。构建知识库的第一步是数据摄取，即将各种来源的数据（如文档、数据库等）整合到系统中.20 接下来是对数据进行清洗和预处理，去除噪声、标准化格式。为了提高检索效率，通常需要将文本数据分割成更小的块（chunking）.20 然后，使用向量嵌入技术将这些文本块转换为能够捕捉语义信息的向量表示.20 这些向量嵌入被存储在向量数据库中，并建立索引以实现快速检索.20 一些流行的向量数据库包括 Pinecone、FAISS、ChromaDB、Weaviate、Qdrant 和 Azure Cosmos DB.20 为了进一步提高检索准确性，可以为知识库中的数据添加元数据标签.20 最后，知识库需要持续更新和维护，以确保信息的时效性和准确性.20

分析表明，高质量的知识库是 RAG 系统成功的基石。精细的数据准备、有效的索引策略以及持续的维护更新是确保 RAG 系统能够提供准确和相关答案的关键要素。向量数据库的多样性为开发者提供了根据应用需求选择合适工具的灵活性。

RAG 系统评估指标与基准的进展

评估 RAG 系统的性能是一个复杂的过程，因为需要同时考虑生成响应的质量和事实准确性.24 传统的自然语言处理评估指标可能无法完全捕捉 RAG 系统的细微之处。因此，研究人员提出了新的评估指标和基准。例如，RAGEval 框架引入了完整性、幻觉和不相关性等指标.24 RAG 三元组则侧重于评估上下文相关性、真实性和答案相关性.25 为了评估 RAG 在长上下文和长格式生成方面的能力，开发了 LONG2RAG 等基准.15 评估检索准确性（精确率、召回率、MRR、MAP）和生成质量（ROUGE、BLEU、BertScore、LLM 作为裁判）仍然很重要.26 此外，利用 LLM 本身作为裁判来评估 RAG 性能的各个方面也成为一种趋势.25 总的来说，需要全面且针对特定场景的评估数据集和指标.24

分析表明，RAG 系统评估领域的进步反映了对该技术复杂性的更深理解。新的评估指标和基准的出现，以及利用 LLM 进行自动评估的方法，有助于更全面地衡量 RAG 系统的性能，并推动该领域的进一步发展。

RAG 在各行业的尖端应用

RAG 技术在各个行业展现出广泛的应用潜力：

- **医疗保健：** 增强临床决策支持，提高响应准确性并减少偏差，实现个性化医疗和精准医疗，简化管理效率，加强患者互动.18
- **金融：** 自动化客户支持，风险分析与管理，研究与投资组合管理，法规遵从与审计支持.32
- **教育：** 个性化学习体验，实时辅助，多样化内容生成，智能辅导系统.38
- **客户支持：** 提供快速准确的响应，处理复杂查询，个性化协助.3
- **法律：** 法律文件分析，案例法研究，合规性监控.3

分析表明，RAG 在不同行业的广泛应用突显了其解决现实世界问题的多功能性和潜力，这些问题需要访问和利用大量信息。从医疗保健到金融再到教育，RAG 正在被各个领域采用，以提高效率、准确性和个性化水平。这些具体的应用案例展示了 RAG 在解决行业特定挑战方面的实际益处。

RAG 开发的最新工具和框架

RAG 的发展离不开各种工具和框架的支持。目前市面上存在许多开源和商业的 RAG 框架，例如 LangChain、LlamaIndex、Haystack、RAGatouille 和 EmbedChain.48 这些框架提供了数据摄取、索引、检索和提示工程等功能。它们还与各种 LLM 提供商（OpenAI、Anthropic、Google、Hugging Face）和向量数据库（Pinecone、ChromaDB、Weaviate、FAISS、Qdrant）集成。此外，还涌现出一些云端 RAG 服务和平台，如 OpenAI Assistants API、Azure AI Search、Amazon Bedrock Knowledge Bases 和 Google Cloud Vertex AI Vector Search.49 新兴的工具和平台包括 RAGFlow、LLMWare.ai、Puppy Agent 和 Neurite.48

以下表格总结了一些流行的向量数据库，它们常用于 RAG 系统：

名称	开源？	托管服务？	价格模型	支持的向量长度	支持的距离度量	索引技术（例如 HNSW、IVF）	元数据过
Pinecone	否	是	按量付费	高	余弦、欧几里得、点积等	HNSW、IVF	是
FAISS	是	否	免费	高	L2、内积、余弦等	LSH、IVF、HNSW、PQ 等	是
ChromaDB	是	否	免费	中	余弦、点积、L2 等	HNSW	是
Weaviate	是	是	分层定价	高	余弦、点积、欧几里得等	HNSW、倒排索引	是
Qdrant	是	是	分层定价	高	余弦、点积、欧几里得等	HNSW、标量量化	是
Azure Cosmos DB	否	是	按量付费	高	余弦、欧几里得、内积等	DiskANN、HNSW	是

以下表格列出了一些关键的 RAG 框架及其主要特点：

框架名称	开源？	主要特点（数据摄取、检索、增强、生成）	LLM 集成	向量数据库集成	常用案例	学习曲线
LangChain	是	模块化设计，支持多种数据源，灵活的提示工程，内存管理	OpenAI、Anthropic、Google、Hugging Face 等	Chroma、Pinecone、FAISS、Weaviate、Qdrant 等	问答系统、聊天机器人、代码生成、文档摘要	中等
LlamaIndex	是	专注于数据索引和检索，支持结构化和非结构化数据，自适应分块	OpenAI、Hugging Face 等	Chroma、Pinecone、FAISS、Weaviate、Qdrant 等	知识助手、文档问答、语义搜索	简单
Haystack	是	检索器-阅读器流水线，混合搜索（BM25、密集向量、神经检索），预构建流水线	Hugging Face Transformers 等	Elasticsearch、OpenSearch、FAISS 等	搜索系统、问答系统、文档检索	中等
RAGatouille	是	轻量级，专注于结合预训练语言模型和高效检索技术	Hugging Face Transformers	FAISS	问答系统、信息检索	简单
EmbedChain	是	简化嵌入管理和集成过程，专注于嵌入的创建和部署	OpenAI、Hugging Face 等	ChromaDB、Pinecone、Weaviate 等	语义搜索、推荐系统	简单

分析表明，RAG 工具和框架的日益普及为开发者提供了丰富的选择，可以根据其具体需求和技术栈选择合适的工具。云端服务的出现进一步简化了 RAG 系统的部署和管理，使得更多组织能够利用这项强大的技术。

RAG 的未来趋势与发展前景

RAG 技术正处于快速发展阶段，未来将呈现出以下几个重要趋势：

- **实时 RAG：** 实现动态检索最新信息，将实时数据流集成到 RAG 模型中.17
- **知识图谱和图嵌入的集成：** 利用知识图谱和图嵌入增强上下文理解.17
- **多模态 RAG 的进步：** 扩展 RAG 以处理文本、图像、音频和视频等多种数据格式.17
- **自查询 RAG 模型的发展：** 使 AI 系统能够自行优化搜索查询，提高查询精度.17
- **RAG 即服务和设备端 AI 实现：** 推动 RAG 技术的普及和更广泛的应用.17
- **解决偏见、确保数据隐私和安全以及维护数据完整性：** 关注 RAG 系统负责任的开发和部署.17
- **检索机制的持续改进：** 包括自适应检索和多阶段检索等技术.17
- **缓存增强生成 (CAG) 等替代范式的出现：** 在某些场景下作为 RAG 的高效替代方案.53
- **强调评估和提高 RAG 系统的可信度和鲁棒性：** 确保 RAG 技术的可靠性和有效性.15

分析表明，RAG 的未来发展将更加注重实时性、多模态能力和智能化。解决伦理问题和提高系统可靠性将是关键的研究方向。同时，新的架构范式的出现也为 RAG 技术的进一步创新提供了可能性。

结论：RAG 的持续进步与巨大潜力

总而言之，检索增强生成 (RAG) 代表了人工智能领域的一项重大进步，它通过将大型语言模型与外部知识相结合，显著提升了生成内容的准确性和相关性。本报告探讨了 RAG 的基本原理、架构创新、知识库构建与维护策略、评估指标的进展、在各行业的尖端应用、最新的工具和框架以及未来的发展趋势。分析表明，RAG 技术正在朝着更专业化、智能化和多模态的方向发展，并在医疗保健、金融、教育和客户支持等多个领域展现出巨大的应用潜力。随着研究的不断深入和技术的持续创新，RAG 有望在未来的 AI 领域发挥越来越重要的作用。

Works cited

1. blogs.nvidia.com, accessed April 8, 2025, https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/#:~:text=So%2C%20What%20Is%20Retrieval%2DAugmented,gap%20in%20how%20LLMs%20work.

2. What is Retrieval-Augmented Generation (RAG)? | Google Cloud, accessed April 8, 2025, https://cloud.google.com/use-cases/retrieval-augmented-generation
3. What is RAG? - Retrieval-Augmented Generation AI Explained - AWS, accessed April 8, 2025, https://aws.amazon.com/what-is/retrieval-augmented-generation/
4. Retrieval-augmented generation - Wikipedia, accessed April 8, 2025, https://en.wikipedia.org/wiki/Retrieval-augmented_generation
5. What is retrieval-augmented generation? - Red Hat, accessed April 8, 2025, https://www.redhat.com/en/topics/ai/what-is-retrieval-augmented-generation
6. What Is Retrieval-Augmented Generation aka RAG - NVIDIA Blog, accessed April 8, 2025, https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/
7. What is RAG (Retrieval Augmented Generation)? - IBM, accessed April 8, 2025, https://www.ibm.com/think/topics/retrieval-augmented-generation
8. What is Retrieval Augmented Generation (RAG)? | A Comprehensive RAG Guide - Elastic, accessed April 8, 2025, https://www.elastic.co/what-is/retrieval-augmented-generation
9. Retrieval Augmented Generation (RAG) - Pinecone, accessed April 8, 2025, https://www.pinecone.io/learn/retrieval-augmented-generation/
10. What is retrieval-augmented generation (RAG)? - IBM Research, accessed April 8, 2025, https://research.ibm.com/blog/retrieval-augmented-generation-RAG
11. What is retrieval-augmented generation (RAG)? - McKinsey & Company, accessed April 8, 2025, https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-retrieval-augmented-generation-rag
12. RAG, or Retrieval Augmented Generation: Revolutionizing AI in 2025 - Glean, accessed April 8, 2025, https://www.glean.com/blog/rag-retrieval-augmented-generation
13. Accepted Main Conference Papers - ACL 2024, accessed April 8, 2025, https://2024.aclweb.org/program/main_conference_papers/
14. What Is RAG Architecture? A New Approach to LLMs - Cohere, accessed April 8, 2025, https://cohere.com/blog/rag-architecture
15. Top 10 RAG Papers from January 2025 - Athina AI Hub, accessed April 8, 2025, https://hub.athina.ai/top-10-rag-papers-from-january-2025-2/
16. The Rise and Evolution of RAG in 2024 A Year in Review | RAGFlow, accessed April 8, 2025, https://ragflow.io/blog/the-rise-and-evolution-of-rag-in-2024-a-year-in-review
17. Trends in Active Retrieval Augmented Generation: 2025 and Beyond, accessed April 8, 2025, https://www.signitysolutions.com/blog/trends-in-active-retrieval-augmented-generation
18. What Are the Future Trends in RAG for 2025 and Beyond? - Chitika, accessed April 8, 2025, https://www.chitika.com/future-trends-in-retrieval-augmented-generation-what-to-expect-in-2025-and-beyond/
19. arxiv.org, accessed April 8, 2025, https://arxiv.org/abs/2407.08223
20. Building a Knowledge Base for RAG: A Step-by-Step Guide | by Arushi Aggarwal | Medium, accessed April 8, 2025, https://medium.com/@arushiagg04/building-a-knowledge-base-for-rag-a-step-by-step-guide-c3afbccf3700
21. Strategies For Turning Mediocre Enterprise Content Into RAG-Ready Knowledge Bases, accessed April 8, 2025, https://blog.gopenai.com/strategies-for-turning-mediocre-enterprise-content-into-rag-ready-knowledge-bases-f83237a224b1
22. Optimizing RAG with Knowledge Base Maintenance - HumanFirst, accessed April 8, 2025, https://www.humanfirst.ai/blog/optimizing-rag-with-knowledge-base-maintenance
23. How to Build a Local RAG Knowledge Base in 2025 | PuppyAgent, accessed April 8, 2025, https://www.puppyagent.com/blog/How-to-Build-a-Local-RAG-Knowledge-Base-in-2025
24. arxiv.org, accessed April 8, 2025, https://arxiv.org/pdf/2408.01262
25. Benchmarking LLM-as-a-Judge for the RAG Triad Metrics - Snowflake, accessed April 8, 2025, https://www.snowflake.com/en/engineering-blog/benchmarking-LLM-as-a-judge-RAG-triad-metrics/

26. Evaluating RAG performance: Metrics and benchmarks - Maxim AI, accessed April 8, 2025, https://www.getmaxim.ai/blog/rag-evaluation-metrics/
27. How Retrieval-Augmented Generation (RAG) Supports Healthcare AI Initiatives, accessed April 8, 2025, https://www.makebot.ai/blog-en/how-retrieval-augmented-generation-rag-supports-healthcare-ai-initiatives
28. How Does Retrieval-Augmented Generation (RAG) Support Healthcare AI Initiatives?, accessed April 8, 2025, https://healthtechmagazine.net/article/2025/01/retrieval-augmented-generation-support-healthcare-ai-perfcon
29. Developing and orchestrating generative AI solutions for healthcare - AWS Documentation, accessed April 8, 2025, https://docs.aws.amazon.com/prescriptive-guidance/latest/rag-healthcare-use-cases/development.html
30. Enhancing Healthcare Diagnostics with Retrieval-Augmented Generation (RAG): Leveraging LangChain, ChromaDB, and Hugging Face | by Sanghvirajit | Mar, 2025, accessed April 8, 2025, https://sanghvirajit.medium.com/enhancing-healthcare-diagnostics-with-retrieval-augmented-generation-rag-leveraging-langchain-2befddfe5859?source=rss-------1
31. Bridging AI and Healthcare: A Scoping Review of Retrieval-Augmented Generation—Ethics, Bias, Transparency, Improvements, and Applications | medRxiv, accessed April 8, 2025, https://www.medrxiv.org/content/10.1101/2025.04.01.25325033v1
32. RAG in 2025: Smarter Retrieval and Real-Time Responses - DataForest, accessed April 8, 2025, https://dataforest.ai/blog/rag-in-2025-smarter-retrieval-and-real-time-responses
33. Optimizing RAG Pipelines in Financial Services: Advanced Strategies from Fitch Group | by ODSC - Open Data Science | Feb, 2025, accessed April 8, 2025, https://odsc.medium.com/optimizing-rag-pipelines-in-financial-services-advanced-strategies-from-fitch-group-cc3e81684817
34. RAG in Financial Services - Signity Software Solutions, accessed April 8, 2025, https://www.signitysolutions.com/blog/rag-in-financial-services
35. AI in Finance: The Promise and Risks of RAG - Lumenova AI, accessed April 8, 2025, https://www.lumenova.ai/blog/ai-finance-retrieval-augmented-generation/
36. 2024 Generative AI Market: Paving the Way for AI in 2025, accessed April 8, 2025, https://www.allganize.ai/en/blog/2024-generative-ai-market-paving-the-way-for-ai-in-2025
37. Building an AI-Powered Financial Advisory Assistant Using Generative AI and RAG | by Kumud Sharma | Mar, 2025 | Medium, accessed April 8, 2025, https://medium.com/@kumud.sharma.0206/building-an-ai-powered-financial-advisory-assistant-using-generative-ai-and-rag-794591196ded
38. The Impact of Generative AI and RAG on Personalized Learning - MAKEBOT.AI, accessed April 8, 2025, https://www.makebot.ai/blog-en/the-impact-of-generative-ai-and-rag-on-personalized-learning
39. RAG Use Case: Unlocking the Potential of Retrieval-Augmented Generation, accessed April 8, 2025, https://www.novusasi.com/blog/rag-use-cases-unlocking-the-potential-of-retrieval-augmented-generation
40. Introduction to RAG - Data Science Institute - The University of Arizona, accessed April 8, 2025, https://datascience.arizona.edu/events/introduction-rag
41. Planning for the RAG Application - State of Michigan, accessed April 8, 2025, https://www.michigan.gov/mde/services/school-performance-supports/statewide-mi-excel/planning-for-the-rag-application
42. Retrieval-Augmented Generation (RAG): The Definitive Guide [2025] - Chitika, accessed April 8, 2025, https://www.chitika.com/retrieval-augmented-generation-rag-the-definitive-guide-2025/
43. Mastering RAG: Enhancing AI Applications with Retrieval-Augmented Generation | by ODSC, accessed April 8, 2025, https://odsc.medium.com/mastering-rag-enhancing-ai-applications-with-retrieval-augmented-generation-846b2bcd8985
44. Retrieval Augmented Generation Explained - Dataiku blog, accessed April 8, 2025, https://blog.dataiku.com/retrieval-augmented-generation-explained
45. 5 key features and benefits of retrieval augmented generation (RAG ..., accessed April 8, 2025, https://www.microsoft.com/en-us/microsoft-cloud/blog/2025/02/13/5-key-features-and-benefits-of-retrieval-augmented-generation-rag/
46. Latest Developments in Retrieval-Augmented Generation - CelerData, accessed April 8, 2025, https://celerdata.com/glossary/latest-developments-in-retrieval-augmented-generation

47. Advanced RAG: Architecture, Techniques, Applications and Use Cases and Development, accessed April 8, 2025, https://www.leewayhertz.com/advanced-rag/
48. Top AI RAG Tools for 2025 - K2view, accessed April 8, 2025, https://www.k2view.com/blog/ai-rag-tools/
49. Compare the Top 7 RAG Frameworks in 2025 - Pathway, accessed April 8, 2025, https://pathway.com/rag-frameworks
50. Top 5 RAG Tools to Kickstart your Generative AI Journey - Analytics Vidhya, accessed April 8, 2025, https://www.analyticsvidhya.com/blog/2024/05/rag-tools/
51. Top 10 Open-Source RAG Frameworks you need!! - DEV Community, accessed April 8, 2025, https://dev.to/rohan_sharma/top-10-open-source-rag-frameworks-you-need-3fhe
52. Top RAG Frameworks Powering the Next Generation of AI Applications - Medium, accessed April 8, 2025, https://medium.com/@skphd/top-rag-frameworks-powering-the-next-generation-of-ai-applications-c83c06bf8d44
53. Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks, accessed April 8, 2025, https://arxiv.org/html/2412.15605v2

其他参考：

https://arxiv.org/html/2503.10677v2

Open AI Deep Research

Gemini new