良性循环：AI驱动的向量搜索与向量搜索增强的人工智能

Jiuqi Wei*Oceanbase*，蚂蚁集团北京，中国
weijiuqi.wjq@antgroup.com

Quanqing Xu*Oceanbase*，蚂蚁集团杭州，中国
xuquanqing.xqq@oceanbase.com

Chuanhui Yang*
Oceanbase，蚂蚁集团北京，中国 ri
zhao.ych@oceanbase.com

arXiv:2603.09347v1 [cs.DB] 10 Mar 2026

摘要——**现代人工智能与向量搜索正在迅速融合，形成智能信息系统领域一个前景广阔的研究前沿。一方面，人工智能的进步显著提升了向量搜索的语义准确性和效率，包括学习型索引结构、自适应剪枝策略和自动化参数调优。另一方面，强大的向量搜索技术催生了新的人工智能范式，尤其是检索增强生成（RAG），它有效缓解了大语言模型（LLM）面临的知识过时和幻觉等挑战。这种相互促进形成了一个“良性循环”：人工智能为向量搜索注入智能与自适应优化能力，而向量搜索则反过来扩展人工智能在知识整合和上下文感知生成方面的能力。本教程全面概述了这一交汇领域的最新研究与进展。我们首先探讨整合向量搜索与人工智能的基础背景和动机。随后，我们深入讨论人工智能如何在向量搜索流水线的各个环节赋能向量搜索（AI4VS）。接着，我们研究向量搜索如何赋能人工智能（VS4AI），重点关注将动态外部知识源整合到大语言模型生成过程中的 RAG 框架。此外，我们还分析充分释放向量搜索与人工智能之间“良性循环”潜力的端到端协同优化策略。最后，我们指出这一新兴领域面临的关键挑战和未来研究机遇。**

索引词—**向量搜索、人工智能、大语言模型、检索增强生成、索引**

I. 引言

过去十年，信息科学经历了两场变革：向量搜索赋予机器类人的语义理解能力，使其能够检索和推荐信息；现代人工智能，尤其是大语言模型的兴起，则赋予机器前所未有的生成和推理能力。这些进展共同从根本上重塑了人类与海量信息资源交互的方式。

如今，向量搜索与人工智能已不再是彼此平行的领域，而是日益融合，形成了前沿研究方向。一方面，人工智能的进步赋予向量搜索智能化能力和自适应优化能力，大幅提升了搜索准确性和效率。学习式索引结构、自适应剪枝策略和自动化参数调优等方面已取得显著进展。另一方面，强大的向量搜索增强了人工智能整合知识和生成上下文感知内容的能力。这催生了检索增强生成（RAG）等新范式，有效缓解了大语言模型（LLM）面临的知识过时和幻觉等挑战。这种相互促进的关系形成了一个“良性循环”：人工智能推动向量搜索（AI4VS）发展，而向量搜索反过来增强人工智能（VS4AI）。本教程将系统分析这种双向关系，阐述人工智能如何优化向量搜索流程，以及反之亦然。

教程概述。本教程全面概述向量搜索与人工智能交叉领域的最新研究与进展。图 1 展示了这一交叉领域的架构。本教程时长为**1.5 小时**，分为以下五个部分：

1) **背景与动机（25 分钟）**。本部分介绍向量搜索和生成式人工智能的基础知识，并通过分析二者各自面临的挑战，阐释向量搜索为何需要人工智能，反之亦然。2) **AI驱动的向量搜索（25 分钟）**。本部分探讨人工智能如何在向量搜索流水线的各个环节赋能向量搜索（AI4VS），包括学习型索引结构、自适应剪枝策略和自动化参数调优。3)**向量搜索增强人工智能（25 分钟）**。本部分研究向量搜索如何赋能人工智能（VS4AI），重点关注检索增强生成框架的演进：从朴素检索增强生成到高级检索增强生成，最终发展为模块化检索增强生成。4) **闭环：端到端优化（10 分钟）**。本部分分析端到端协同优化策略，以充分释放向量搜索与人工智能之间“良性循环”的潜力。5)**挑战与机遇（5 分钟）**。本部分重点介绍这一新兴领域的关键挑战和未来研究方向。

目标受众。本教程面向对向量搜索、人工智能及两者交集感兴趣的学生、研究人员和开发者。本教程提供全面且独立的入门介绍，无需具备特定的先备知识。

*Chuanhui Yang 是本文的通讯作者。

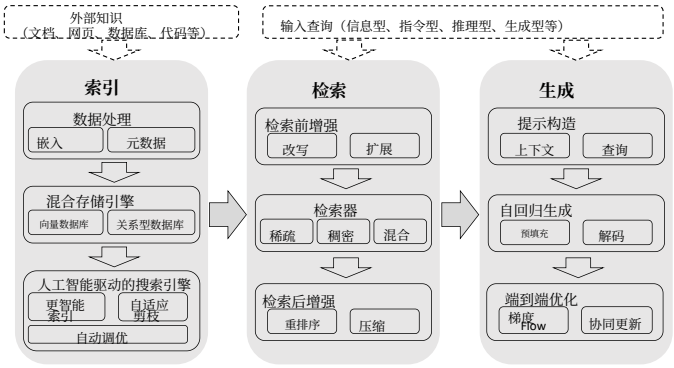


图1. AI驱动的向量搜索（AI4VS）、向量搜索增强的人工智能（VS4AI）及其端到端优化的架构。

II. 教程大纲

A. 背景与动机

我们首先介绍向量搜索和生成式人工智能的基础知识，并通过分析两者各自面临的挑战，阐明为什么向量搜索需要人工智能，反之亦然。

1) 向量搜索概述：向量搜索可以从三个核心方面进行系统介绍：基础问题、技术方法和实际系统。

问题。我们首先介绍向量搜索背后的基本问题：近似最近邻搜索（ANNS）、最大余弦相似度搜索（MCSS）和最大内积搜索（MIPS），以及它们之间的数学关系和转换方法。

方法。接下来，我们对最先进的向量搜索方法进行分类，将其划分为局部敏感哈希方法 [1]–[3], 子空间碰撞方法 [4], 向量量化方法 [5]–[7], 树方法 [8]–[10], 图方法 [11]–[13], 以及混合方法 [14]–[16],，并系统分析各自的优势与局限。

系统。最后，我们介绍现代向量数据库的架构，例如 FAISS [17], Milvus [18],ScaNN [19], 和 VSAG [20], ，这些系统结合了多种向量搜索技术，以提升可扩展性并优化延迟。

2) 向量搜索面临的挑战：传统向量搜索面临一系列挑战，而现代人工智能技术能够有效缓解这些挑战（解决方案见第 II-B 节）。

索引不准确。传统索引结构依赖专家经验和启发式设计规则，难以适应不同的数据分布，也无法有效处理边界区域的划分。因此，如何为数百万乃至数十亿条数据构建更智能的索引，使索引结构能够根据数据分布动态调整，成为一项关键挑战。

低效查询。传统向量搜索方法对所有查询都采用固定的索引遍历策略，无法根据当前查询的特征（例如难易程度）动态调整计算资源，导致查询开销浪费。因此，如何设计自适应剪枝机制，智能确定搜索路径和深度，以避免搜索过程中不必要的遍历，已成为一项关键挑战。

非最优参数。向量搜索系统涉及许多参数，从索引构建到查询应答都需要仔细调优。传统的参数调优依赖专家手动试错，成本高、效果不理想，也无法适应动态数据变化。因此，实现自动化参数调优成为一项关键挑战。

*3)*生成式人工智能概述：对生成式人工智能进行系统性介绍，可以围绕三项关键技术展开：作为基础模型的Transformer架构、赋予其上下文理解能力的自注意力机制，以及规模化实现的大语言模型。

Transformer。我们首先介绍 Transformer架构 [21], 它是现代生成式人工智能的基础构件。我们将讨论关键的架构变体，包括仅编码器 [22], 、仅解码器 [23], 和编码器-解码器 [24], ，并说明每种变体对应的应用任务。

自注意力。接下来，我们将深入探讨自注意力机制 [21]的数学公式及计算范式。我们将详细说明查询-键-值机制的运作方式，以及它如何为 Transformer 提供上下文理解能力。

大语言模型。最后，我们介绍大语言模型，将其视为Transformer架构的规模化实现。讨论还涵盖了大语言模型的发展历程，以及当前最先进的开源 [25]–[27] 和闭源 [23], [28]模型。

4) 生成式人工智能的挑战：大语言模型面临若干挑战，而向量搜索技术可以有效应对这些挑战（解决方案见第 II-C 节）。

知识过时。大语言模型基于静态语料库进行训练，无法了解训练完成后发生的近期事件或不断变化的信息。因此，其知识很快就会过时，也可能无法涵盖特定领域的信息或最新事实。

幻觉。面对知识缺口或含糊不清的查询时，大语言模型容易生成看似合理、实则不符合事实甚至完全捏造的内容，因为它们依赖的是内部统计模式，而非经过核实的证据。

成本高昂。将更多世界知识编码进模型需要扩展参数规模，这会导致训练和推理成本急剧上升；而扩大模型规模并不能保证提升事实召回率。

缺乏特异性。通用大语言模型基于公开语料库训练，在专业领域（例如法律、医疗、企业专属知识）或用户特定上下文方面表现不佳，若不进行成本高昂的微调，便难以应对这些需求。

B. AI驱动的向量搜索（AI4VS）

我们探讨如何利用人工智能技术在智能和自适应优化方面的强大能力，解决第II-A2节中概述的向量搜索挑战。

1) 更智能的索引：学习哈希方法利用深度神经网络将高维向量转换为紧凑的哈希码，同时保留原始空间与编码空间之间的语义相似性，从而实现高效

的压缩域搜索 [29]–[31]。学习划分方法制定数据驱动的划分策略，以适应底层数据分布和簇边界，从而提高整体索引精度 [32]–[35]。向量量化方法使用机器学习算法，通过最小化重构误差或保留邻域结构，将高维向量压缩为紧凑的量化码，从而实现高效且精度可控的近似搜索 [5], [6], [36]。

2) 自适应剪枝：提前终止策略利用模型估算检索真实 k 近邻或达到声明式召回率所需访问的最少向量数或簇数，从而实现查询感知搜索并降低计算成本 [37], [38]。分区剪枝利用模型预测有效搜索所需的、针对特定查询的分区数量，动态将遍历范围限制在相关子集中，并尽量减少不必要的分区访问 [32], [39]。路由剪枝在概率建模框架下重新构建图路由问题，通过学习高效的路由策略，引导在邻近图上搜索近邻 [40]。精化剪枝运用学习得到的正交变换，并在精化阶段逐步收紧距离界；每当候选项的下界超过当前最佳距离阈值时，即可实时进行候选剪枝 [41]。

3) 自动化参数调优：基于搜索的方法通过定义庞大的配置空间，并以实证方式评估大量候选参数集，从而确定最优配置 [42], [43]。尽管这种方法全面，但计算成本可能很高。基于学习的方法借助机器学习模型（例如贝叶斯优化、分类器或预训练元模型），智能推荐有望达到目标召回率阈值的配置，或直接预测各种参数设置下的查询性能 [44]–[46]。这一范式显著减少了对穷举式人工测试的需求。混合方法将基于搜索的方法的稳健性与基于学习的方法的高效性相结合，为更具适应性且能够自我优化的向量数据库铺平道路 [20]。

C. 向量搜索增强人工智能（*VS4AI*）

我们探讨向量搜索如何利用其在语义索引和高效检索方面的强大能力，解决第 II-A4 节中概述的大语言模型面临的挑战。

向量搜索通过检索增强生成，系统性地解决大语言模型的核心挑战；检索增强生成会有策略地将动态外部知识源整合到生成过程中 [47]–[48]。检索增强生成框架已从朴素检索增强生成发展到 AdvancedRAG，最终演进为 ModularRAG [49]。

*1)*朴素检索增强生成（检索增强生成 ***1.0***）：朴素检索增强生成阶段的核心创新在于建立一个稳定、固定且线性的框架，以便将外部的非参数知识注入生成过程 [50]–[52]。朴素检索增强生成遵循一条基础流水线，其中包括索引、检索和生成 [49]。

- 索引：在这一准备阶段，外部知识语料库（例如文档、网页）被划分为文本块。随后，这些文本块通过嵌入模型编码为向量表示，从而构建一个能够实现高效相似度比较的索引。
- 检索：收到用户查询后，系统会以类似方式将其编码为向量。随后，系统在预构建索引上执行相似度搜索，以识别并检索最相关的 top-k 个文本块。
- 生成：最后，将检索上下文与原始查询拼接成提示词，并输入大语言模型。随后，大语言模型根据提供的上下文综合生成流畅连贯的回答。

朴素检索增强生成流程成功证明，利用检索增强大语言模型能够显著提高事实准确性和可靠性 [53]。

2) 高级检索增强生成（***RAG 2.0***）： 尽管朴素检索增强生成流程取得了成功，但它存在一个结构性弱点：极易受到检索质量的影响。高级检索增强生成框架着重优化流程中的各个组成部分，尤其是检索前、检索和检索后模块，以最大限度地提升输入上下文的质量。

- 检索前增强：这类工作旨在改进索引结构或查询本身，从而获得更好的搜索结果。索引优化会改变检索增强生成系统中知识的组织方式，以实现更精准的信息检索。主要方法包括保留语义连贯性的自适应分块策略 [54], 、丰富文档表示的策略性元数据附加 [55], [56],, 以及能够捕捉复杂语义关系的知识图谱增强 [57], [58]。查询优化着重改进用户查询，使其更好地表达信息需求，并与检索系统的特性相匹配。这包括查询改写，即由大语言模型将原始查询改写为更有效的版本 [59]–[61], ；以及查询扩展，即生成多个相关查询，以检索范围更广的相关文档 [62], [63]。
- 检索增强：混合检索 [64], 结合了稀疏方法和稠密方法，利用两者互补的优势，显著增强了检索增强生成在应对复杂现实场景查询时的鲁棒性。稀疏检索 [65]确保精确的词汇匹配，但难以处理语义变化；稠密检索 [51] 能够捕捉语义相似度，却可能遗漏关键的关键词匹配。融合函数根据稀疏检索器和稠密检索器的异构分数 [66], [67]确定最终的统一排序。多向量模型是混合检索范式中的一项重要演进 [68], [69]。这些模型使用多个细粒度向量而非单个稠密嵌入来表示文档和查询，从而实现更细致的语义匹配 [70]–[72]。
- 检索后增强：在初步检索到

候选段落后，会进行关键的后处理阶段，以在生成前优化上下文，并使其与查询有效融合。重排序 可优化检索增强生成中的检索结果，确保只有最相关的文档会传递给生成器 [73], [74]。根据语义相关性和查询对齐情况重新排列候选内容，可以有效减少噪声、提升事实性，并提高生成回复的整体质量。上下文压缩 通过将检索到的文本精简为更短、信息更密集的输入 [75]–[77]，缓解大语言模型上下文窗口有限的问题。这一步能提高效率、降低成本，并使检索增强生成系统能够扩展到更大的语料库和更长的检索上下文。

*3)*模块化检索增强生成（检索增强生成 ***3.0***）：模块化检索增强生成代表了检索增强生成的发展方向，它超越了固定的线性流程（朴素检索增强生成）和组件级优化（高级检索增强生成），为 RAG 框架引入了架构灵活性。这种模块化设计支持动态组合工作流并进行专项优化，通过灵活集成和替换组件，提升复杂任务的处理表现。

- 核心模块：模块化检索增强生成开发了专用组件，用以取代线性流程。路由器 动态选择处理路径 [78]。记忆 模块保留上下文历史，以支持多轮交互 [79]。预测器 生成假设性上下文，以改进检索 [80]。搜索 模块支持针对不同数据源进行定向查询 [81]。

- 高级工作流模式：除了引入新模块，模块化检索增强生成还探索如何将这些模块定制组合为非线性工作流。迭代式模式支持多步推理循环 [82]。演示-搜索-预测 模式利用示例驱动检索，并主动生成上下文 [83]。条件式 模式使系统能够根据查询需求自适应地调用模块 [84]。

D. 闭环：端到端优化

尽管检索增强生成已展现出将向量搜索与大语言模型相结合的强大能力，但其检索器与生成器仍以不同目标分别优化。检索器最大化查询-文档相似度，而生成器则关注文本似然。这种目标错位导致协调效果欠佳。

我们分析了近期关于端到端优化的研究，这类研究在统一框架内联合训练检索器与生成器，以充分释放“良性循环”的潜力。

- 可微检索：与使用固定不可微索引的传统检索不同，可微检索使生成器损失的梯度能够反向传播并优化检索器 [85], [86]。

- 检索器-生成器联合训练：一些方法并不让检索过程变得可微，而是协调检索器和生成器的更新，使二者都朝着同一个最终任务目标（例如问答准确率）进行优化 [87]。检索器可能仍然选择离散的 top-k 文档，梯度也可能无法直接通过检索过程传播。相反，模型会通过间接信号交替或联合更新。

E. 挑战与机遇

自演化索引。大多数学习索引方法遵循静态范式：在固定数据集上训练一次后即部署，不再进一步适应。然而，现实世界的向量数据库本质上是动态的：随着上游模型更新，嵌入会发生漂移，数据分布也在持续变化。因此，开发能够在无需完全重新训练的情况下增量调整分区边界、哈希函数或量化码的持续索引学习器变得至关重要。此类自适应系统可在数据流下保持检索性能，同时保留与现有索引结构的向后兼容性。

检索记忆化。检索增强生成系统面临严峻挑战，包括重复检索周期导致的高延迟、冗余向量搜索带来的计算成本上升，以及对语义相似查询的响应不一致。这些局限凸显了对有效 RAG 缓存机制的迫切需求，该机制通过存储并智能复用先前检索或处理的结果来缓解这些低效问题。通过避免不必要的检索和生成步骤，RAG 缓存可大幅缩短响应时间并降低计算开销，同时提升输出的稳定性与可复现性。

自主执行。模块化 RAG 框架因依赖预定义工作流且缺乏动态推理而面临根本性局限。它们无法根据中间结果重新规划，也无法自我评估信息充分性。这要求其演进为智能体式 RAG，后者引入动态规划、持续自我监控以及主动的工具使用。通过作为能够制定并执行多步策略的自主智能体，智能体式 RAG 将系统从无状态执行器转变为认知助手，使其能够借助策略推理与自适应学习处理复杂且模糊的任务。

III. 主讲人

Jiuqi Wei 是蚂蚁集团 Oceanbase Lab 的研究员。他在中国科学院计算技术研究所获得博士学位。他的研究兴趣包括向量数据库、混合检索和数据管理。他近期的研究成果 [3], [4], [88]–[97] 主要聚焦于近似最近邻搜索和数据治理。

Quanqing Xu 是蚂蚁集团 Oceanbase Lab 的技术总监。他的研究兴趣主要包括分布式数据库系统和存储系统 [98]–[110]。他是 IET 会士、中国计算机学会杰出会员，以及 ACM 和 IEEE 高级会员。

Chuanhui Yang 是蚂蚁集团 OceanBase 的首席技术官。他的研究主要聚焦于数据库、分布式系统和人工智能基础设施 [99]–[109]。作为创始成员之一，他领导了 OceanBase 早期的架构设计与技术研发，使 OceanBase 从零起步，最终在蚂蚁集团全面落地。

人工智能生成内容声明

在本项研究过程中，作者使用大语言模型（Gemini 和 ChatGPT）辅助润色文本和写作。使用这些工具后，作者对内容进行了必要的审阅和编辑，并对发表内容承担全部责任。

参考文献

[1] Y. Tian、X. Zhao 和 X. Zhou，《DB-LSH 2.0：采用基于查询的动态分桶的局部敏感哈希》，IEEE知识与数据工程汇刊，2023。[2] B. Zheng、Z. Xi、L. Weng、N. Q. V. Hung、H. Liu 和 C. S. Jensen，《PM-LSH：一种用于高维近似最近邻搜索的快速、精确的局部敏感哈希框架》，VLDB基金会会刊，第 13 卷，第 5 期，第 643–655 页，2020。[3] J. Wei、B. Peng、X. Lee 和 T. Palpanas，《Det-lsh：一种采用动态编码树的局部敏感哈希方案，用于近似最近邻搜索》，VLDB基金会会刊，第 17 卷，第 9 期，第 2241–2254 页，2024。[4] J. Wei、X. Lee、Z. Liao、T. Palpanas 和 B. Peng，《子空间碰撞：一种高效、精确的高维近似最近邻搜索框架》，ACM数据管理会刊，第 3 卷，第 1 期，第 1–29 页，2025。[5] H. Jegou、M. Douze 和 C. Schmid，《用于最近邻搜索的乘积量化》，IEEE模式分析与机器智能汇刊，第 33 卷，第 1 期，第 117–128 页，2010。[6] T. Ge、K. He、Q. Ke 和 J. Sun，《优化乘积量化》，IEEE模式分析与机器智能汇刊，第 36 卷，第 4 期，第 744–755 页，2013。[7] J. Gao 和 C. Long，《Rabitq：以理论误差界量化高维向量，用于近似最近邻搜索》，ACM数据管理会刊，第 2 卷，第 3 期，第 1–27 页，2024。[8] E. Bernhardsson，《C++/Python 中经优化的近似最近邻搜索库，专为节省内存及从磁盘加载/保存而设计》，2015。[在线]。地址：https://github.com/spotify/annoy [9] S. Dasgupta 和 Y. Freund，《随机投影树与低维流形》，载于 ACM计算理论研讨会第四十届年会论文集，2008，第 537–546 页。[10] K. Zoumpatianos、S. Idreos 和 T. Palpanas，《ADS：自适应数据序列索引》，VLDB期刊，第 25 卷，第 843–866 页，2016。[11] J. V. Munoz、M. A. Gonc,alves、Z. Dias 和 R. d. S. Torres，《基于层次聚类图的大规模近似最近邻搜索》，模式识别，第 96 卷，文章编号 106970，2019。[12] C. Fu、C. Xiang、C. Wang 和 D. Cai，《利用导航扩展图实现快速近似最近邻搜索》，VLDB基金会会刊，第 12 卷，第 5 期，第 461–474 页，2019。[13] Y. A. Malkov 和 D. A. Yashunin，《利用分层可导航小世界图进行高效、稳健的近似最近邻搜索》，IEEE模式分析与机器智能汇刊，第 42 卷，第 4 期，第 824–836 页，2018。[14] Q. Chen、H. Wang、M. Li、G. Ren、S. Li、J. Zhu、J. Li、C. Liu、L. Zhang 和 J. Wang，《Sptag：用于快速近似最近邻搜索的库》，2018。[15] I. Azizi、K. Echihabi 和 T. Palpanas，《Elpis：用于可扩展数据科学的基于图的相似度搜索》，VLDB基金会会刊，第 16 卷，第 6 期，第 1548–1559 页，2023。[16] Y. Gou、J. Gao、Y. Xu 和 C. Long，《Symphonyqg：面向量化与图在近似最近邻搜索中的和谐集成》，ACM数据管理会刊，第 3 卷，第 1 期，第 1–26 页，2025。[17] M. Douze、A. Guzhva、C. Deng、J. Johnson、G. Szilvasy、P.-E. Mazar´e、M. Lomeli、L. Hosseini 和 H. J´egou，《FAISS 库》，IEEE大数据汇刊，2025。[18] J. Wang、X. Yi、R. Guo、H. Jin、P. Xu、S. Li、X. Wang、X. Guo、C. Li、X. Xu 等，《Milvus：专为向量数据管理打造的系统》，载于 2021 年国际数据管理会议论文集，2021，第 2614–2627 页。[19] R. Guo、P. Sun、E. Lindgren、Q. Geng、D. Simcha、F. Chern 和 S. Kumar，《通过各向异性向量量化加速大规模推理》，载于国际机器学习会议。PMLR，2020，第 3887–3896 页。

[20] X. Zhong, H. Li, J. Jin, M. Yang, D. Chu, X. Wang, Z. Shen, W. Jia, G. Gu, Y. Xie 等，“VSAG：一种用于基于图的近似最近邻搜索的优化搜索框架”，arXiv 预印本 arXiv:2503.17911，2025。[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser 和 I. Polosukhin，“Attention Is All You Need”，神经信息处理系统进展，第 30 卷，2017。[22] J. Devlin, M.-W. Chang, K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向 Transformer 预训练”，载于《2019年计算语言学协会北美分会会议：人类语言技术论文集》第1卷（长篇与短篇论文），2019，第4171–4186页。[23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat 等，“GPT-4技术报告”，arXiv 预印本 arXiv:2303.08774，2023。[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li 和 P. J. Liu，“探索统一文本到文本 Transformer 的迁移学习极限”，机器学习研究期刊，第21卷，第140期，第1–67页，2020。[25] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang 等，“通义千问技术报告”，arXiv 预印本 arXiv:2309.16609，2023。[26] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan 等，“DeepSeek-V3技术报告”，arXiv 预印本 arXiv:2412.19437，2024。[27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale 等，“Llama 2：开放基础模型与微调聊天模型”，arXiv 预印本 arXiv:2307.09288，2023。[28] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican 等，“Gemini：一系列能力出众的多模态模型”，arXiv 预印本 arXiv:2312.11805，2023。[29] M. Li, Y. Zhang, Y. Sun, W. Wang, I. W. Tsang 和 X. Lin，“基于学习函数的 I/O 高效近似最近邻搜索”，载于《2020年IEEE第36届国际数据工程会议（ICDE）》。IEEE，2020，第289–300页。[30] Q.-Y. Jiang 和 W.-J. Li，“非对称深度监督哈希”，载于《AAAI 人工智能会议论文集》，第32卷，第1期，2018。[31] H. Lai, Y. Pan, Y. Liu 和 S. Yan，“利用深度神经网络同时进行特征学习与哈希编码”，载于《IEEE计算机视觉与模式识别会议论文集》，2015，第3270–3278页。[32] X. Zeng, L. Deng, P. Chen, X. Chen, H. Su 和 K. Zheng，“Lira：一种面向大规模近似最近邻搜索的基于学习的查询感知分区框架”，载于《2025年ACM Web 会议论文集》，2025，第2729–2741页。[33] G. Gupta, T. Medini, A. Shrivastava 和 A. J. Smola，“Bliss：一种通过迭代式重新分区构建的十亿级索引”，载于《第28届ACM SIGKDD 知识发现与数据挖掘会议论文集》，2022，第486–495页。[34] W. Li, C. Feng, D. Lian, Y. Xie, H. Liu, Y. Ge 和 E. Chen，“学习用于大规模向量检索的平衡树索引”，载于《第29届ACM SIGKDD 知识发现与数据挖掘会议论文集》，2023，第1353–1362页。[35] Y. Dong, P. Indyk, I. P. Razenshteyn 和 T. Wagner，“学习用于最近邻搜索的空间分区”，《国际学习表征会议》，2020。[36] Y. Kalantidis 和 Y. Avrithis，“用于近似最近邻搜索的局部优化乘积量化”，载于《IEEE计算机视觉与模式识别会议论文集》，2014，第2321–2328页。[37] C. Li, M. Zhang, D. G. Andersen 和 Y. He，“通过学习式自适应提前终止改进近似最近邻搜索”，载于《2020年ACM SIGMOD 数据管理国际会议论文集》，2020，第2539–2554页。[38] M. Chatzakis, Y. Papakonstantinou 和 T. Palpanas，“Darth：通过提前终止实现声明式召回率的近似最近邻搜索”，SIGMOD，2026。[39] B. Zheng, Z. Yue, Q. Hu, X. Yi, X. Luan, C. Xie, X. Zhou 和 C. S. Jensen，“面向高维近似近邻搜索的学习式探测基数估计”，载于《2023年IEEE第39届国际数据工程会议（ICDE）》。IEEE，2023，第3209–3221页。

[40] D. Baranchuk, D. Persiyanov, A. Sinitsin 和 A. Babenko，“学习在相似性图中进行路由”，载于《国际机器学习会议》。PMLR，2019，第475–484页。[41] V. Ramani, A. Schlomer, A. Nayar, P. Karras, S. Ranu 和 J. M. Patel，“Panorama：快速近邻搜索”，arXiv 预印本 arXiv:2510.00566，2025。[42] W. Zhou, Y. Jiang, Y. Liu, X. Qiao, H. Zhang, H. Li 和 J. Cui，“自动调优分层可导航小世界图的构建参数”，SSRN 4925468。[43] M. Wang, X. Xu, Q. Yue 和 Y. Wang，“基于图的近似最近邻搜索：全面综述与实验比较”，VLDB基金会会刊，第14卷，第11期，第1964–1978页，2021。[44] T. Yang, W. Hu, W. Peng, Y. Li, J. Li, G. Wang 和 X. Liu，“Vdtuner：向量数据管理系统的自动化性能调优”，载于《2024年IEEE第四十届国际数据工程会议（ICDE）》。IEEE，2024，第4357–4369页。[45] H. Duan, Y. Song, B. Yao 和 A. Liang，“Pgtuner：一种用于邻近图自动化与可迁移配置调优的高效框架”，ACM数据管理会刊，第3卷，第4期，第1–27页，2025。[46] R. S. Oyamada, L. C. Shimomura, S. B. Junior 和 D. S. Kaster，“迈向邻近图自动配置：一种基于元学习的方法”，载于《欧洲数据库与信息系统进展会议》。施普林格，2020，第93–107页。[47] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua 和 Q. Li，“RAG 与大语言模型融合综述：迈向检索增强大语言模型”，载于《第30届ACM SIGKDD 知识发现与数据挖掘会议论文集》，2024，第6491–6501页。[48] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang 和 B. Cui，“面向人工智能生成内容的检索增强生成：综述”，arXiv 预印本 arXiv:2402.19473，2024。[49] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang 和 H. Wang，“面向大语言模型的检索增强生成：综述”，arXiv 预印本 arXiv:2312.10997，第2卷，第1期，2023。[50] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. K¨uttler, M. Lewis, W.-t. Yih, T. Rockt¨aschel 等，“面向知识密集型自然语言处理任务的检索增强生成”，神经信息处理系统进展，第33卷，第9459–9474页，2020。[51] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen 和 W.-t. Yih，“面向开放域问答的密集段落检索”，载于 EMNLP（1），2020，第6769–6781页。[52] G. Izacard 和 E. Grave，“将知识从阅读器蒸馏到检索器以实现问答”，arXiv 预印本 arXiv:2012.04584，2020。[53] A. J. Oche, A. G. Folashade, T. Ghosal 和 A. Biswas，“检索增强生成（RAG）关键系统的系统综述：进展、差距与未来方向”，arXiv 预印本 arXiv:2507.18910，2025。[54] C. Merola 和 J. Singh，“重构上下文：评估用于检索增强生成的高级分块策略”，载于《知识增强信息检索国际研讨会》。施普林格，2025，第3–18页。[55] S. Gollapudi, N. Karia, V. Sivashankar, R. Krishnaswamy, N. Begwani, S. Raz, Y. Lin, Y. Zhang, N. Mahapatro, P. Srinivasan 等，“Filtered-DiskANN：带过滤器的近似最近邻搜索图算法”，载于《2023年ACM Web 会议论文集》，2023，第3406–3416页。[56] Y. Cai, J. Shi, Y. Chen 和 W. Zheng，“导航标签与向量：一种用于过滤式近似最近邻搜索的统一方法”，ACM数据管理会刊，第2卷，第6期，第1–27页，2024。[57] Y. Wang, N. Lipka, R. A. Rossi, A. Siu, R. Zhang 和 T. Derr，“用于多文档问答的知识图谱提示”，载于《AAAI 人工智能会议论文集》，第38卷，第17期，2024，第19 206–19 214页。[58] L. Liang, Z. Bo, Z. Gui, Z. Zhu, L. Zhong, P. Zhao, M. Sun, Z. Zhang, J. Zhou, W. Chen 等，“KAG：通过知识增强生成提升专业领域中的大语言模型”，载于《ACM网络会议伴随论文集》，2025，第334–343页。[59] X. Ma, Y. Gong, P. He, N. Duan 等，“检索增强大语言模型中的查询改写”，载于《2023年自然语言处理实证方法会议》，2023。

[60] W. Peng, G. Li, Y. Jiang, Z. Wang, D. Ou, X. Zeng, D. Xu, T. Xu, and E. Chen，《基于大语言模型的淘宝搜索长尾查询改写》，载于《ACM Web 会议 2024 伴随论文集》，2024 年，第 20–28 页。[61] L. Gao, X. Ma, J. Lin, and J. Callan，《无需相关性标签的精准零样本稠密检索》，载于《计算语言学协会第 61 届年会论文集（第 1 卷：长篇论文）》，2023 年，第 1762–1777 页。[62] D. Zhou, N. Sch¨arli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schu- urmans, C. Cui, O. Bousquet, Q. Le 等，《由少至多提示使大语言模型能够进行复杂推理》，arXiv 预印本 arXiv:2205.10625，2022。[63]R. Jagerman, H. Zhuang, Z. Qin, X. Wang, and M. Bendersky，《通过提示大语言模型进行查询扩展》，arXiv 预印本 arXiv:2305.03653，2023。[64] K. Sawarkar, A. Mangal, and S. R. Solanki，《混合式 RAG：通过语义搜索和基于混合查询的检索器提升 RAG（检索器增强生成）的准确性》，载于《2024 年 IEEE 第七届多媒体信息处理与检索国际会议（MIPR）》。IEEE，2024 年，第 155–161 页。[65] S. Robertson, H. Zaragoza 等，《概率相关性框架：BM25 及其后续发展》，《信息检索基础与趋势》，第 3 卷，第 4 期，第 333–389 页，2009 年。[66] S. Bruch, S. Gai, and A. Ingber，《混合检索融合函数分析》，《ACM信息系统汇刊》，第 42 卷，第 1 期，第 1–35 页，2023 年。[67] Z. Rackauckas，《RAG-Fusion：检索增强生成的新思路》，arXiv 预印本 arXiv:2402.03367，2024。[68] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia，《ColBERTv2：通过轻量级后期交互实现有效且高效的检索》，载于《2022 年计算语言学协会北美分会会议：人类语言技术论文集》，2022 年，第 3715–3734 页。[69] J. Engels, B. Coleman, V. Lakshman, and A. Shrivastava，《DESSERT：一种用于向量集查询的高效向量集搜索算法》，《神经信息处理系统进展》，第 36 卷，第 67 972–67 992 页，2023 年。[70] Z. Bian, M. L. Yiu, and B. Tang，《IGP：通过邻近图索引实现高效多向量检索》，载于《第 48 届 ACM SIGIR 信息检索研究与开发国际会议论文集》，2025 年，第 2524–2533 页。[71] R. Jayaram, L. Dhulipala, M. Hadian, J. D. Lee, and V. Mirrokni，《MUVERA：通过固定维度编码实现多向量检索》，《神经信息处理系统进展》，第 37 卷，第 101 042–101 073 页，2024 年。[72] F. M. Nardini, C. Rulli, and R. Venturini，《利用位向量进行高效多向量稠密检索》，载于《欧洲信息检索会议》。施普林格，2024 年，第 3–17 页。[73] J. Dong, B. Fatemi, B. Perozzi, L. F. Yang, and A. Tsitsulin，《别忘了建立连接！通过基于图的重排序改进 RAG》，arXiv 预印本 arXiv:2405.18414，2024。[74] Y. Yu, W. Ping, Z. Liu, B. Wang, J. You, C. Zhang, M. Shoeybi, and B. Catanzaro，《RankRAG：在大语言模型中统一上下文排序与检索增强生成》，《神经信息处理系统进展》，第 37 卷，第 121 156–121 184 页，2024 年。[75] X. Cheng, X. Wang, X. Zhang, T. Ge, S.-Q. Chen, F. Wei, H. Zhang, and D. Zhao，《xRAG：使用单个词元进行检索增强生成的极端上下文压缩》，《神经信息处理系统进展》，第 37 卷，第 109 487–109 516 页，2024 年。[76] D. Rau, S. Wang, H. D´ejean, and S. Clinchant，《用于 RAG 高效答案生成的上下文嵌入》，arXiv 预印本 arXiv:2407.09252，2024。[77] Y. Fang, T. Sun, Y. Shi, and X. Gu，《AttentionRAG：检索增强生成中的注意力引导上下文剪枝》，arXiv 预印本 arXiv:2503.10720，2025。[78] X. Li, E. Nie, and S. Liang，《从分类到生成：跨语言检索增强上下文学习的启示》，arXiv 预印本 arXiv:2311.06595，2023。[79] X. Cheng, D. Luo, X. Chen, L. Liu, D. Zhao, and R. Yan，《提升自我：通过自记忆实现检索增强文本生成》，《神经信息处理系统进展》，第 36 卷，第 43 799 页，2023 年，43 780 页。

[80] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang, 《生成而非检索：大语言模型是强大的上下文生成器》，arXiv 预印本 arXiv:2209.10063，2022。[81] X. Wang, Q. Yang, Y. Qiu, J. Liang, Q. He, Z. Gu, Y. Xiao, and W. Wang, 《Knowledgpt：通过检索和知识库存储访问增强大语言模型》，arXiv 预印本 arXiv:2308.11761，2023。[82] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, 《通过迭代检索—生成协同增强检索增强大语言模型》，载于《计算语言学协会会议论文集：EMNLP 2023》，2023 年，第 9248–9274 页。[83] O. Khattab, K. Santhanam, X. L. Li, D. Hall, P. Liang, C. Potts, and M. Zaharia, 《Demonstrate-search-predict：组合检索与语言模型以实现知识密集型自然语言处理》，arXiv 预印本 arXiv:2212.14024，2022。[84] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, 《Self-rag：通过自我反思学习检索、生成和评判》，载于《国际学习表征会议》，2024 年。[85] 匿名, 《D-RAG：用于知识图谱问答的可微检索增强生成》，提交至 ACL Rolling Review - 2025 年 5 月，2025 年，审稿中。 [在线]。链接：https://openreview.net/forum?id

= 0 3 D vilzHmI C. Zhao, Y.

[86] Jiang, Y. Qiu, H. Zhang, and W.-Y. Yang, 《通过生成式语言建模实现可微检索增强，用于电商查询意图分类》，载于《第 32 届 ACM 信息与知识管理国际会议论文集》，2023 年，第 4445–4449 页。[87] D. Singh, S. Reddy, W. Hamilton, C. Dyer, and D. Yogatama, 《面向开放域问答的多文档阅读器与检索器端到端训练》，《神经信息处理系统进展》，第 34 卷，第 25 968–25 981 页，2021 年。[88] J. Wei, X. Lee, Y. Fu, Y. Li, and B. Peng, 《自主掌控数据：当数据与应用解耦时实现数据互操作的去中心化方案》，《万维网》，第 28 卷，第 3 期，第 1–39 页，2025 年。[89] J. Wei, Y. Li, Y. Fu, Y. Zhang, and X. Li, 《数据互操作架构（DIA）：解耦数据与应用，归还数据所有权》，载于《2023 年 IEEE 第 47 届计算机、软件与应用会议（COMPSAC）》。IEEE，2023 年，第 438–447 页。[90] Y. Li, J. Wei, Z. Fei, Y. Fu, and X. Lee, 《Disauth：用于保护与应用解耦数据的基于 DNS 的安全授权框架》，《计算机网络》，第 254 卷，第 110774 页，2024 年。[91] Z. Zhuang, X. Lee, J. Wei, Y. Fu, and A. Zhang, 《CBCMS：跨境数据传输合规管理系统》，载于 BigData，2024 年。[92] Y. Fu, X. Lee, J. Wei, Y. Li, and B. Peng, 《保障互联网骨干安全：基于区块链和激励机制的 DNS 缓存投毒防御架构》，《计算机网络》，2024 年。[93] Y. Fu, J. Wei, Y. Li, B. Peng, and X. Li, 《Ti-DNS：基于区块链的可信激励型 DNS 解析架构》，载于 TrustCom，2023 年。[94] Z. Fei, Y. Li, J. Wei, Y. Fu, B. Peng, and X. Li, 《Flexauth：具有灵活委托功能的去中心化授权系统》，载于 TrustCom，2023 年。[95] A. Zhang, X. Lee, Z. Zhuang, J. Wei, Y. Fu, and B. Peng, 《Polaris：通过可验证身份和基于策略的授权实现跨域访问控制》，arXiv 预印本 arXiv:2511.22017，2025 年。[96] Z. Zhuang, X. Lee, A. Zhang, J. Wei, Y. Fu, and B. Peng, 《面向全球软件系统多司法辖区合规的结构化策略建模与上下文感知生成》，《Information and Software Technology》，第 108041 页，2026 年。[97] C. Song, Z. Peng, J. Wei, and C. Yang, 《CSAttention：用于加速大语言模型推理的质心评分注意力》，2026 年。[98] Z. Yang, C. Qian, X. Teng, F. Kong, F. Han, Q. Xu, and D. Hu, 《Lcl+：为 OceanBase 构建的基于锁链长度的分布式死锁检测与解决服务》，《ACM Transactions on Computer Systems》，第 43 卷，第 4 期，第 1–33 页，2025 年。[99] C. Fang, C. Qian, Q. Yang, Z. Wang, Z. Yang, F. Kong, Q. Xu, H. Cao, F. Han, and C. Yang, 《Malt：OceanBase 大型事务管理框架》，载于《2025 年国际数据管理会议伴随论文集》，2025 年，第 378–390 页。[100] D. Hu, Q. Xu, and C. Yang, 《现代存储架构上的 OLTP 引擎》，载于《2025 年国际数据管理会议伴随论文集》，2025 年，第 805–812 页。[101] Q. Xu, W. Sun, C. Yang, J. Liu, Z. Wei, F. Han, L. Wang, and X. Zhai, 《OceanBase 单元化：构建新一代在线地图应用”，载于*2025*年*IEEE*第*41*届数据工程国际会议（*ICDE*）。IEEE，2025年，第4183–4196页。[102] Q. Luo、Q. Xu和C. Yang，“如何回答安全与隐私SQL查询？”，载于*2025*年*IEEE*第*41*届数据工程国际会议（*ICDE*）。IEEE，2025年，第4486–4491页。[103] Q. Luo、J. Yu、W. Dong、Q. Xu、C. Yang和K. Yi，“RM2：在洗牌差分隐私下高效回答计数查询”，载于*ACM*数据管理会刊，第3卷，第3期，第1–24页，2025年。[104] Z. Yang、C. Yang、F. Han、M. Zhuang、B. Yang、Z. Yang、X. Cheng、Y. Zhao、W. Shi、H. Xi 等，“OceanBase：每分钟处理7.07亿次事务的分布式关系数据库系统”，载于*VLDB*基金会会刊，2022年。[105] Z. Yang、Q. Xu、S. Gao、C. Yang、G. Wang、Y. Zhao、F. Kong、H. Liu、W. Wang和J. Xiao，“OceanBase Paetica：支持单机和分布式集群的混合共享无一切架构/共享一切架构数据库”，载于*VLDB*基金会会刊，2023年。[106] Q. Zhang、J. Li、H. Zhao、Q. Xu、W. Lu、J. Xiao、F. Han、C. Yang和X. Du，“异构网络中的高效分布式事务处理”，载于*VLDB*基金会会刊，第16卷，第6期，第1372–1385页，2023年。[107] L. Qu、Q. Wang、T. Chen、K. Li、R. Zhang、X. Zhou、Q. Xu、Z. Yang、C. Yang、W. Qian 等，“当前基准是否足以评估分布式事务数据库？”，载于*BenchCouncil*基准、标准与评估汇刊，第2卷，第1期，文章编号100031，2022年。[108] Q. Xu、C. Yang和A. Zhou，“原生分布式数据库：问题、挑战与机遇”，载于*VLDB*基金会会刊，第17卷，第12期，第4217–4220页，2024年。[109] F. Han、H. Liu、B. Chen、D. Jia、J. Zhou、X. Teng、C. Yang、H. Xi、W. Tian、S. Tao 等，“Palf：面向分布式数据库的复制式预写日志”，载于*VLDB*基金会会刊，第17卷，第12期，第3745–3758页，2024年。[110] Z. Yang、C. Qian、X. Teng、F. Kong、F. Han和Q. Xu，“LCL：基于锁链长度的分布式算法，用于死锁检测与解决”，载于*2023*年*IEEE*第*39*届数据工程国际会议（*ICDE*）。IEEE，2023年，第151–163页。