基于英特尔® 技术构建 RAG 模块

深入研究检索增强生成 (Retrieval Augmented Generation, RAG)，该创新方法定义了企业和机构如何利用大语言模型 (LLM) 来发挥其数据的价值。本文将探索若干英特尔® 软硬件构建模块如何帮助优化 RAG 应用，在简化部署和支持扩展的同时，增强其上下文感知能力和实时响应性能。

作者： Eduardo Alvarez，英特尔高级 AI 解决方案工程师 • Sancha Norris，英特尔 AI 产品营销部门

为您的应用量身定制 GenAI

ChatGPT 的面世改变了 AI 的发展格局。企业争相利用这项新技术打造新产品，提高竞争优势和生产力，实现更加经济高效的运营。

生成式 AI（GenAI）模型，如 Grok-1（逾 3,000 亿参数）和 GPT-4（数万亿参数），利用来自互联网等文本来源的海量数据进行训练。这些第三方大语言模型适用于通用用例。然而，企业的大多数用例都需要使用自身的数据来训练和/或增强 AI 模型，这样模型产出的结果才能对业务更有帮助。以下是生成式 AI 在各行各业的应用示例。

消费品和零售	医疗和医药	制造业	媒体和娱乐	金融服务
▪ 虚拟试衣间 ▪ 配送和安装 ▪ 店内产品查找帮助 ▪ 需求预测和库存规划 ▪ 新产品设计	▪ 协助忙碌的一线员工 ▪ 转录和总结医疗记录 ▪ 聊天机器人和回答医学问题 ▪ 通过预测性分析为诊断和治疗方案提供依据	▪ 技术人员专业助手 ▪ 与机器的对话互动 ▪ 规范性和主动性现场服务 ▪ 自然语言故障排除 ▪ 保修状态和文档记录 ▪ 了解流程瓶颈，制定恢复策略	▪ 智能搜索、定制化内容发现 ▪ 标题和文稿起草 ▪ 对内容质量的实时反馈 ▪ 个性化播放列表、新闻摘要、推荐 ▪ 通过观众选择进行互动式故事讲述 ▪ 定向优惠、订阅计划	▪ 发现交易信号，提醒交易员注意脆弱头寸 ▪ 加速做出包销决策 ▪ 优化和重建老旧系统 ▪ 银行和保险模型逆向工程 ▪ 监测潜在的金融犯罪和欺诈 ▪ 根据合规性要求自动执行数据收集 ▪ 从企业披露信息中获取洞察

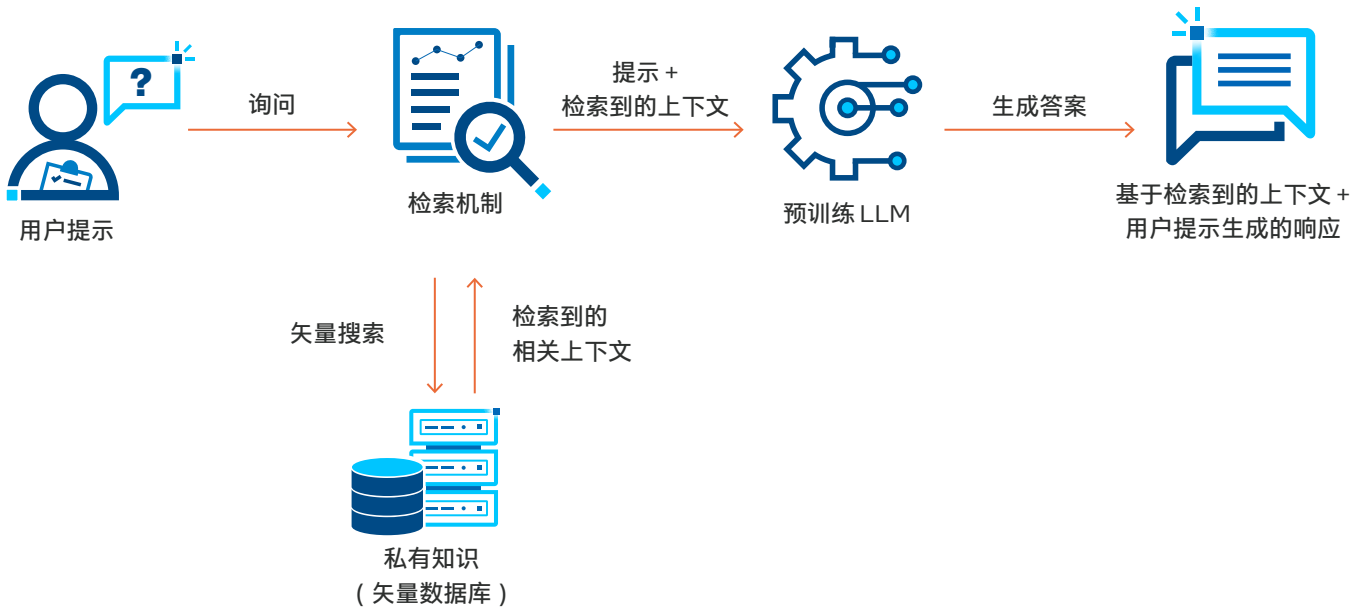
来源：由《麻省理工科技评论》根据"Retail in the Age of Generative AI（生成式AI时代的零售）"、"The Great Unlock: Large Language Models in Manufacturing（大解锁：制造业中的大语言模型）"、"Generative AI Is Everything Everywhere, All at Once（生成式 AI 无处不在、每时每刻都在发生）"和"Large Language Models in Media & Entertainment（媒体和娱乐行业中的大语言模型）"（Databricks，2023 年 4 月至 6 月）中的数据编写。

虽然企业可以用自有的数据对模型进行调优，但重新训练模型需要额外的时间和资源。好在现在有了一种颇受欢迎的技术，即检索增强生成（RAG），它可以利用企业专有的数据来增强开源预训练模型，从而创建特定领域的 LLM，得出针对具体业务的结果。此外，RAG 无需与第三方大型基础模型共享数据，因此能够让企业更好地保护数据安全。

在本指南中，我们将介绍 RAG 如何与英特尔多种优化技术和平台搭配使用，为 GenAI 系统带来出色的价值和性能。

检索增强生成(RAG)是什么?

RAG 技术将动态、依赖查询的数据添加到模型的提示流中，再从存储在矢量数据库中的专有知识库中检索相关数据。提示和检索到的上下文可以丰富模型的输出，从而带来更加相关和准确的结果。因为数据不会被发送给管理模型的第三方，因此，RAG 可让企业在保护数据隐私性和完整性的同时更好地通过 LLM 充分利用数据。RAG 工作流程的关键构成可简单分为四个步骤：用户查询处理、检索、上下文整合和输出生成。下图展示了这一基本流程。



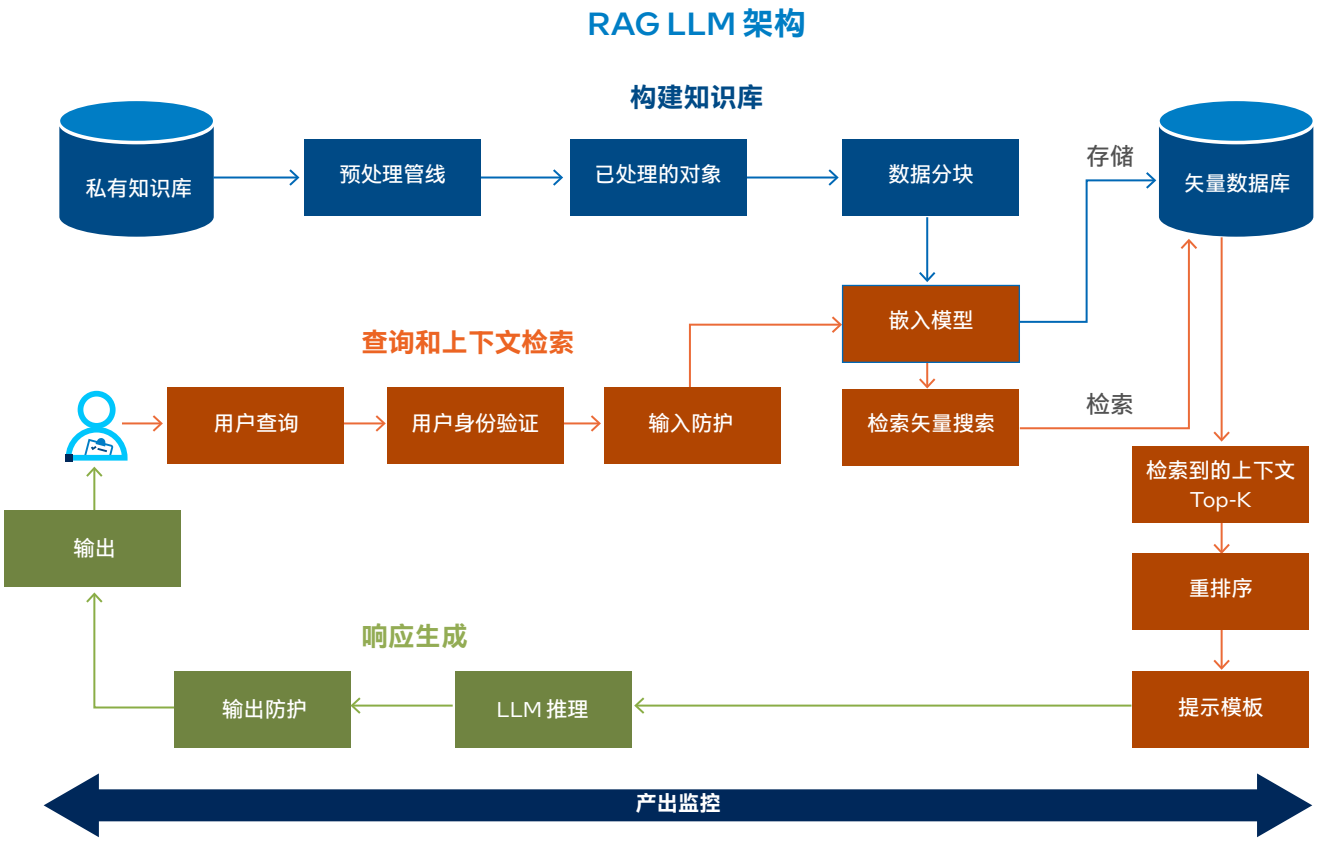
RAG 的实用性不仅限于文本，它还可以极大地改变视频搜索和交互式文档探索的方式，甚至使聊天机器人能够利用 PDF 内容来回答问题。

RAG 的应用过程通常被称为“RAG 管线”，因其从用户提示开始，整个数据处理流程都是一致的。用户提示首先进入关键步骤“检索机制”中。在这一步，相关提示会被转换为矢量嵌入，接着使用矢量搜索在预先构建的矢量数据库（如 PDF、日志、转录文本等）中找到相似的内容。检索到最相关的数据后，RAG 会将其与用户提示整合，然后传送给模型用于推理服务和最终输出生成。这种上下文整合为模型提供了在预训练阶段无法获得的额外信息，使模型能够更好地契合用户的任务或兴趣领域。由于 RAG 无需重新训练或调优模型，因此能够高效地添加数据来为 LLM 提供上下文。

下一节将探讨 RAG 解决方案的架构和堆栈。

标准 RAG 解决方案的架构

下图所示的 RAG 解决方案架构展示了标准 RAG 实施方案的构建模块。RAG 实施流程主要包括 1 构建知识库、2 查询和上下文检索、3 响应生成和 4 跨应用产出监控几个核心部分。



让我们展开谈谈其中几个核心部分：

1 构建知识库：

- **数据收集：** 从基于文本的来源（如转录文本、PDF 和数字化文档）中收集数据建立私有知识库。
- **数据处理管线：** 利用特定 RAG 管线来提取文本、格式化内容以进行处理，并将数据分块成可管理的大小。
- **矢量化：** 通过嵌入模型处理数据块，将文本转换为矢量，可包括用于丰富上下文的元数据。
- **矢量数据库存储：** 将矢量化数据存储在可扩展的矢量数据库中，以便进行高效检索。

2 查询和上下文检索：

- **查询提交：** 用户或子系统通过聊天式界面或 API 调用提交查询，并通过安全服务进行身份验证。
- **查询处理：** 采取输入保护措施来确保安全性和合规性，然后进行查询矢量化。
- **矢量搜索和重排序：** 进行初始矢量搜索以检索相关矢量，然后使用更复杂的模型重排序以优化结果。

3 响应生成：

- **LLM 推理和响应生成：** 将顶层上下文与用户查询结合，再通过预训练或调优的 LLM 进行处理，然后再进行后处理以提升质量和增强安全性。
- **响应交付：** 通过界面将最终响应返回给用户或子系统，确保答案的连贯性和上下文准确性。

4 产出监控：

- **检索性能：** 监控检索过程的时延和准确性，并保留记录以用于审核。
- **重排序的效率：** 跟踪重排序的表现，确保上下文相关性和速度。
- **推理服务质量：** 观察 LLM 推理的时延和质量，维护日志以便审核和改进。
- **安全防护有效性：** 监控输入和输出处理的安全防护（guardrail），确保合规性和内容安全性。

RAG 相关技术

开发 RAG 应用通常会从集成 RAG 框架开始，例如 Haystack、LlamaIndex、LangChain 和英特尔研究院的 fastRAG。这些框架可通过提供优化和集成关键的 AI 工具链来简化开发过程。

我们从知识库构建、查询和上下文检索以及响应生成这三个关键步骤来考量 RAG 工具链。通常，RAG 框架提供涵盖整个工具链的 API。不管是选择使用这些抽象，还是选择利用独立组件，都需要深思熟虑并从工程角度慎重考虑。

英特尔提供的优化技术/方案填补了工具链和硬件之间的缺口，并且在提升这些工具链与英特尔® 至强® 处理器兼容性及功能的同时，增强了跨工具链的性能。这些优化被集成到现有框架中，或者作为附加的扩展进行分发，目的是减少开发人员对大量低级别编程的需求。这种抽象使得开发人员能够利用增强的性能和针对其特定用例量身定制的解决方案，专注于高效构建 RAG 应用。

接下来，本文将对工具链的多个组成部分进行更详细的探讨。

构建知识库 + 上下文检索：

- **集成框架：** Haystack 和 LangChain 作为常见 RAG 框架，为矢量数据库和搜索算法提供了高级抽象，使得开发人员能够在基于 Python 的环境中管理复杂的过程。
- **矢量数据库技术：** Pinecone、Redis 和 Chroma 是支持主流搜索算法的关键矢量数据库解决方案。英特尔研究院提供的可扩展矢量搜索 (Scalable Vector Search, SVS) 技术也很有发展前景，预计将在 2024 年初与各大矢量数据库集成。
- **嵌入和模型可访问性：** 通过 Hugging Face API 进行集成的嵌入模型往往可无缝整合到 RAG 框架中。这大大提升了纳入先进自然语言处理 (NLP) 的简便性。

响应生成：

- **低级别优化**：oneAPI 高性能库可以优化 PyTorch、TensorFlow 和 ONNX 等主流 AI 框架，因此您可以使用熟悉的开源工具，因为它们已针对英特尔® 硬件进行了优化。
- **高级推理优化：** 英特尔® Extension for PyTorch 等扩展添加了高级量化推理技术，可助力提升了大语言模型的性能。

如您所见，RAG 涉及多个相关联的部分，在单一平台（如英特尔® 至强® 处理器）上进行管理可简化配置、部署和维护。

下一节将深入探讨 RAG 应用的复杂性，包括帮助团队实现成功部署的各种考量因素和技术。

加速 RAG 应用与落地

RAG 管线的许多步骤需要耗费大量计算资源，而同时，终端用户又对低时延响应有着较高要求。此外，由于 RAG 经常用于处理机密数据，因此整个管线的安全性都至关重要。英特尔® 技术赋能 RAG 管线，助力提升各个计算平台的安全性能和充分发挥专为特定领域或行业量身定制的生成式 AI 的优势。

计算需求

一般来说，LLM 推理是 RAG 管线计算最密集的阶段，特别是在实时应用环境中。然而，创建初始知识库（处理数据和生成嵌入）对计算的需求同样可能很高（取决于数据的复杂性和体量）。英特尔在通用计算技术、AI 加速器和机密计算方面的进步为应对整个 RAG 管线的计算挑战提供了重要基石，同时还能提高数据隐私和安全性。

和大多数软件应用一样，RAG 也能从专为满足终端用户事务需求而量身定制的可扩展基础设施中受益。随着事务需求的增加，开发人员可能会因计算基础设施负载过重而面临时延增加，且基础设施还会因矢量数据库查询和推理计算而趋于饱和。因此，获得随时可用的计算资源来扩展系统和快速处理新增需求对企业至关重要。另外，实施关键优化以提升诸如嵌入生成、矢量搜索与推理等关键步骤的性能也非常重要。

数据隐私和安全性

- **安全 AI 处理：**英特尔® 软件防护扩展（Intel® Software Guard Extensions，英特尔® SGX）和英特尔® Trust Domain Extensions（英特尔® TDX）在处理过程中在 CPU 内存中进行机密计算和数据加密，提高了数据安全性。这些技术对于处理敏感信息至关重要，有助于利用管线各部分的加密数据创建更安全的 RAG 应用。对于需要在矢量嵌入生成、检索或推理过程中更安全地处理敏感数据的 RAG 应用来说，这是一个重要特性。
- **采取适当防护：**在 RAG 应用中，防护涉及采取措施来管理 LLM 在 RAG 系统内的行为。这包括监控模型的响应、帮助遵守指导原则和最佳实践，以及控制其输出来降低毒性、不公平偏见和隐私泄露的风险。在 RAG 应用中采取防护措施有助于 LLM 得到用户的信任和负责任的运用，同时符合系统的整体目标和要求。

开源优化

嵌入优化

- **量化嵌入模型：**英特尔® 至强® 处理器可以利用量化嵌入模型来优化从文档中生成矢量嵌入的过程。例如，*bge-small-en-v1.5-rag-int8-static* 是一个使用英特尔® Neural Compressor 进行量化的 BAAI/BGE-small-en-v1.5 版本，与 Optimum-Intel 兼容。按照 Massive Text Embedding Benchmark（MTEB）性能指标计算，使用量化模型进行检索和重排序任务时，浮点（FP32）和量化 INT8 版本之间的差异小于 2%，同时提高了吞吐量（见脚注 1 和 3）。

在最近与 Hugging Face 合作进行的一项研究中，我们评估了以每秒文档数为指标达到峰值编码性能所需吞吐量。总体而言，无论模型大小，量化模型在各种批大小下均较基线 bfloat16（BF16）模型取得高达 4 倍的改进。详情请访问：https://huggingface.co/blog/intel-fast-embedding

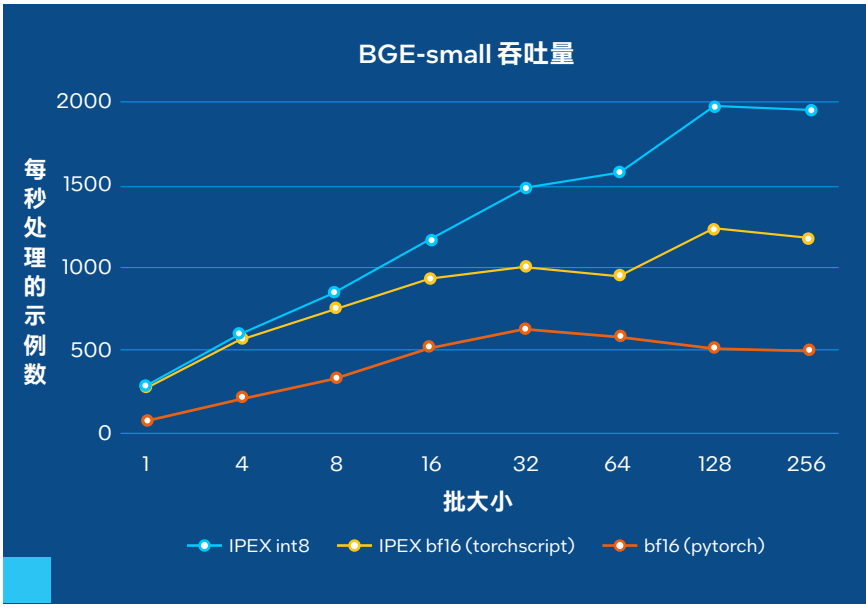


图 1. BGE-small 吞吐量
来源：https://huggingface.co/blog/intel-fast-embedding

矢量搜索优化

- **针对 CPU 优化的工作负载：**在英特尔® 至强® 处理器上，矢量搜索操作得到了高度优化，特别是在第三代及以后的处理器中引入了英特尔® 高级矢量扩展 512（Intel® Advanced Vector Extensions 512，英特尔® AVX-512）之后。英特尔® AVX-512 利用融合乘加（FMA）指令，将乘法和加法合并为一个运算，从而增强了内积计算，这是矢量搜索中的一个基本运算。这一功能减少了计算所需的指令数量，显著提高了吞吐量和性能。
- **可扩展矢量搜索（SVS）：**可扩展矢量搜索（SVS）技术提供快速的矢量搜索能力，可助力优化检索时间并提升整体系统性能。它通过使用局部自适应矢量量化（LVQ）来优化基于图形的相似度搜索，在保持准确性的同时尽可能降低内存带宽要求。其结果是显著减少了距离计算时延，并在吞吐量和内存要求方面获得了更好的表现（如下图所示）。

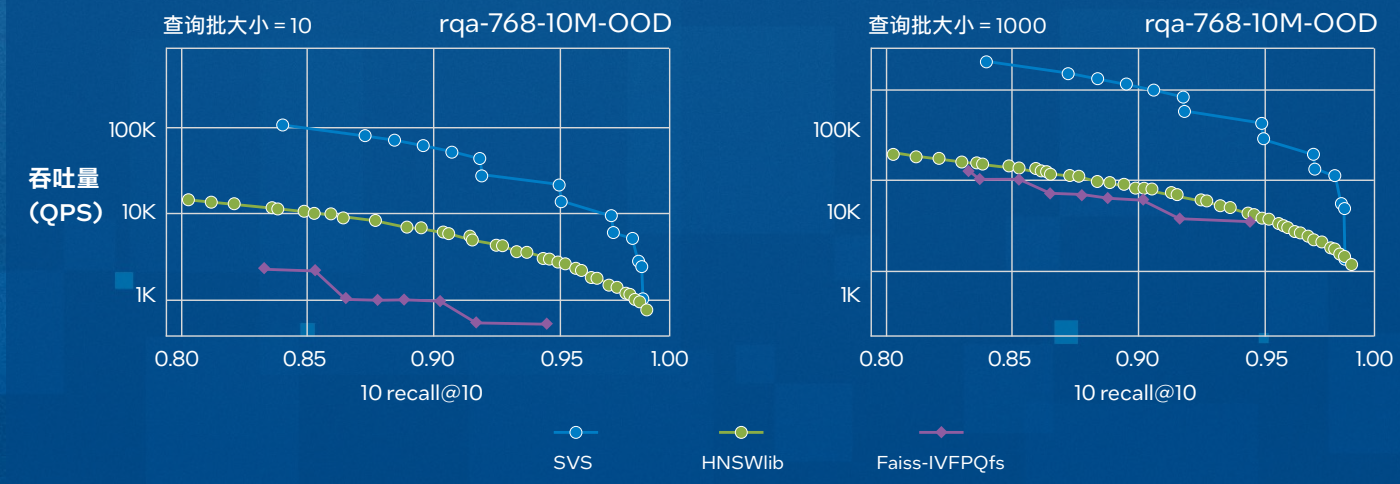
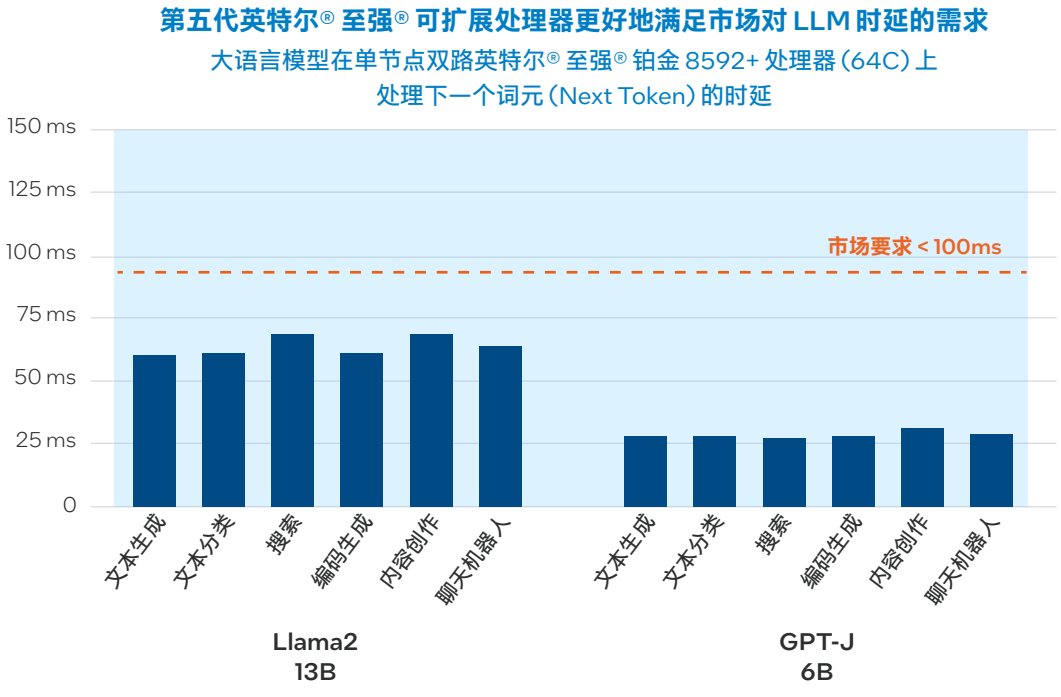


图 2. SVS 与其它被广泛采用的实现方案（HNSWlib 与 Faiss-IVFPQfs）在每秒查询数量（吞吐量）方面的性能对比。该图展示了在 rqa-768-10M-OOD 数据集（由密集通道检索模型 RocketQA[QDLL21] 使用分布外查询生成的 1000 万个768 维嵌入向量）上，QPS 和召回率的关系曲线。（脚注 2 和 3）
来源：https://intellabs.github.io/ScalableVectorSearch/benchs/static/latest.html

推理优化

RAG 主要涉及推理运算，这一过程可由英特尔® 至强® 处理器通过先进的模型压缩技术提供支持。这些技术支持在较低精度（BF16 和 INT8）下进行运算，并且不会造成明显的性能损失。在本节中，我们将简要介绍各种针对推理的优化和机会。

- **英特尔® 高级矩阵扩展（Intel® Advanced Matrix Extensions，英特尔® AMX）：** 第四代和第五代英特尔® 至强® 可扩展处理器内置英特尔® AMX，能够提高矩阵运算的效率并优化内存管理。
- **先进的开源推理优化工具：** 英特尔贡献并扩展了主流深度学习框架，如 PyTorch、TensorFlow、Hugging Face、DeepSpeed 等。对于 RAG 工作流程，英特尔关注的是通过实施量化等模型压缩技术来优化 LLM 的机会。英特尔® Extension for PyTorch 目前提供多种先进的 LLM 量化配方，如 SmoothQuant、仅权重量化和混合精度（FP32/BF16）。下图显示了在双路第五代英特尔® 至强® 平台上运行的 INT8 量化 Llama 2 和 GPT-J 模型的推理时延。



工作负载/配置信息见附页。结果可能不同。

图 3. 基于第五代英特尔® 至强® 可扩展处理器的 Llama 2 13B 和 GPT-J 6B 性能[3]

RAG 在企业中的应用机遇

零售

零售商面临的挑战是向客户推荐符合其多样化和不断变化的偏好的产品。传统的推荐系统可能无法有效地掌握最新趋势或个体客户反馈，导致建议不太贴合实际。

采用基于 RAG 的推荐系统使零售商能够不断整合最新趋势和个体客户反馈，从而得出更个性化的产品建议。该系统通过提供相关、及时和个性化的产品推荐来丰富购物体验，进而助力提高销量并提升客户忠诚度。

了解更多 >

制造业

在制造业中，设备故障导致的意外停机是一个重要的成本驱动因素。传统的预测性维护模型可能会遗漏故障发生前出现的细微异常状况，尤其是历史故障数据有限或缺失的复杂设备的异常状况。

用于预测性维护的基于 RAG 的异常检测系统可以实时分析大量运行数据，并将其与丰富的设备性能知识库进行比对，以在故障发生之前识别出可能存在的问题。这种方法在延长设备使用寿命的同时，尽可能减少了停机时间和维护成本。

了解更多 >

金融服务

由于金融数据和法规不断变化且数量庞大，大规模提供个性化的金融建议面临重重挑战。客户期望能够获得快速、相关且个性化的金融建议，而传统的聊天机器人无法始终准确提供这些建议。

RAG 模型则能够通过动态拉取最新的金融数据和法规来生成个性化的建议，显著增强了金融建议聊天机器人的能力。聊天机器人可以利用庞大的知识库，为客户提供量身定制的投资策略、实时市场洞察和监管建议，从而提高客户满意度和参与度。

了解更多 >

后续行动

英特尔提供一套资源来帮您开始执行实施方案，您可以通过英特尔® Tiber™ Developer Cloud 获取硬件，也可以利用 Google Cloud Platform、Amazon Web Services 和 Microsoft Azure 等各大云服务平台中无处不在的计算资源。对于需要代码示例、演练、培训等内容的开发人员，请访问英特尔® 开发人员专区。

英特尔® Tiber™ Developer Cloud

基于全新英特尔® 至强® 处理器
使用经英特尔优化的软件加速 AI 开发。

基于全新英特尔® 至强® 处理器和其他英特尔® 平台使用经英特尔优化的软件加速 AI 开发。

获取英特尔® 硬件并开始在 Amazon Web Services、Google Cloud Platform 和 Microsoft Azure 等云平台上构建 RAG 应用。

助您基于英特尔® 硬件和软件进行开发的可用官方资源

探索英特尔热门开发领域和资源。

英特尔的 GenAI 开发资源

[1] 性能声明基于双路英特尔® 至强® 铂金 8480+ 处理器，每路 56 个内核。PyTorch 模型使用单路处理器上的 56 个内核进行评估。IPEX/Optimum 设置使用 ipexrun、单路处理器和 22 至 56 个内核进行评估。TCMalloc 在所有运行中都已安装并定义为环境变量。详情请见 www.intel.cn/performanceindex。结果可能不同。

[2] 性能声明基于双路英特尔® 至强® 铂金 8480L 处理器，每路 56 个内核，每路配备 512 GB DDR4 内存，速度为 4800 MT/s，运行 Ubuntu 22.04.12。对于 deep-96-1B 数据集，我们使用具有相同特性的服务器，唯一的区别是每路配备 1 TB DDR4 内存，速度为 4400 MT/s。详情请见 www.intel.cn/performanceindex。结果可能不同。

[3] 实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.cn/PerformanceIndex。性能测试结果基于配置信息中显示的日期进行的测试，且可能并未反映所有公开可用的安全更新。没有任何产品或组件是绝对安全的。具体成本和结果可能不同。英特尔技术可能需要启用硬件、软件或激活服务。© 英特尔公司版权所有。英特尔、英特尔标识以及其他英特尔商标是英特尔公司或其子公司的商标。其他的名称和品牌可能是其他所有者的资产。

0424/SN/MESH/PDF