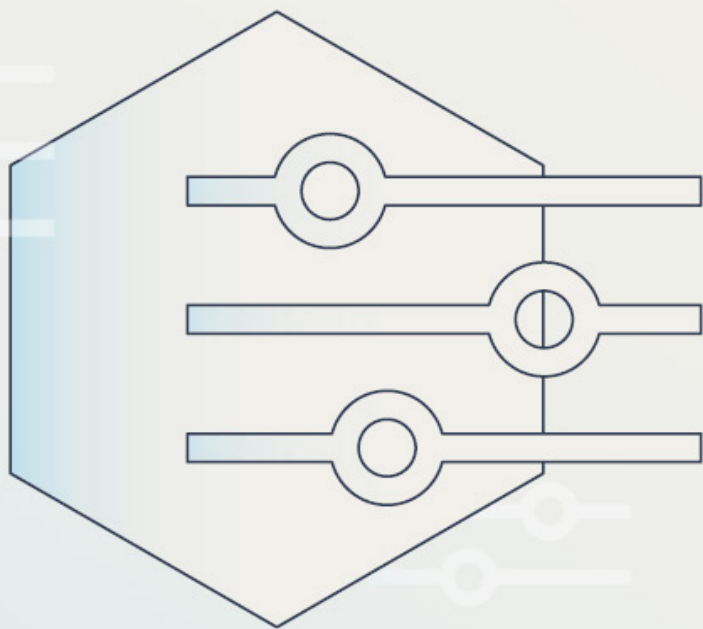
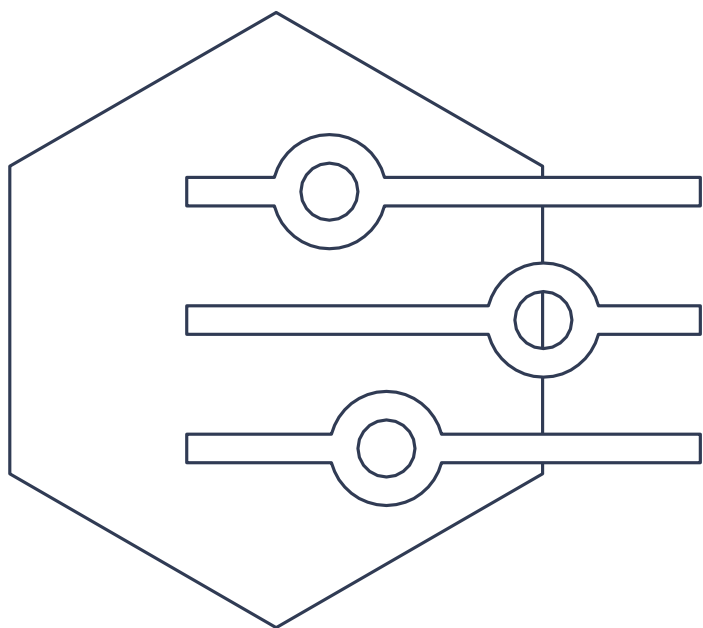


Appen



检索增强生成
(RAG)和人类专
业知识如何优化
AI性能



检索增强生成 (RAG)和人类专 业知识如何优化 AI性能

Appen

版权所有© Appen 2024 – 保留所有权利

目录

i. 检索增强生成(RAG)简介

- a. RAG的最佳用例
- b. RAG的架构
- c. 在RAG中创造卓越：关键决策
- d. 确保RAG输出的质量

ii. RAG优化的数据准备

- a. 导入过程
- b. 查询过程
- c. 持续评估和优化

iii. RAG系统中的人类专业知识

- a. 人类如何改进RAG结果
- b. 人类专业知识在RAG系统中的作用

iv. 自动评估指标：监控RAG的准确性和相关性

- a. 检索指标
- b. 生成指标

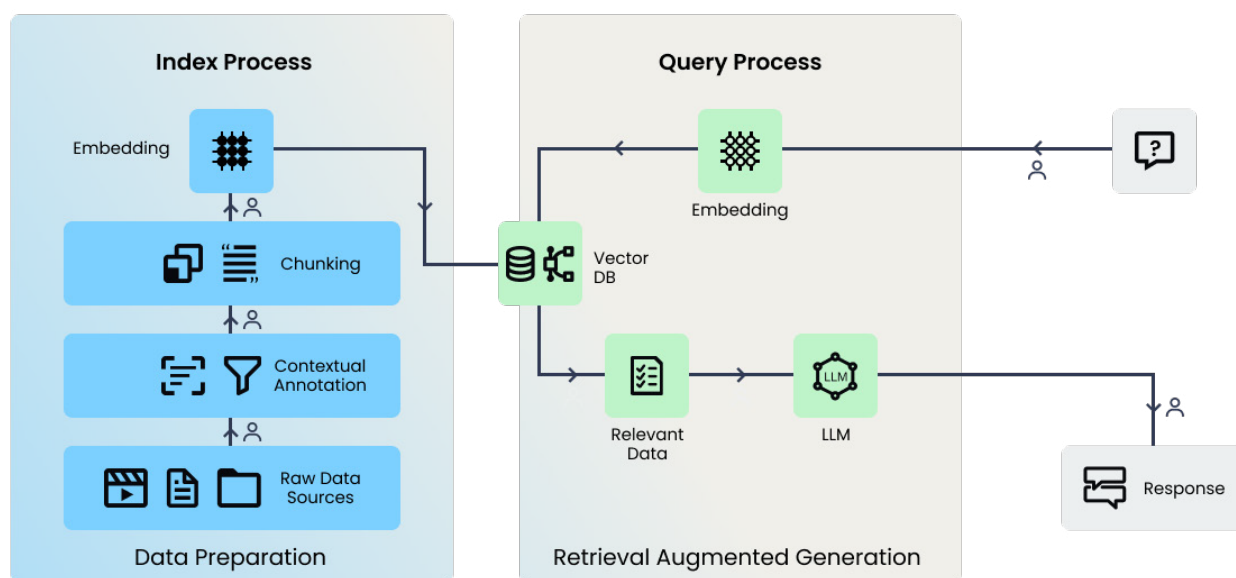
iv. 澳鹏的“Build My RAG”（构建我的RAG）功能

- a. 流程中的步骤
- b. 模板提高效率

v. 结语

检索增强生成(RAG)简介

在生成式AI的发展历程中，[检索增强生成\(RAG\)](#)技术无疑是一项里程碑式的重大进步，既有基于检索的模型准确性又有生成模型创造力，相得益彰。当任务要求精确检索信息以及基于上下文信息做出易于理解的回答时，这种创新的架构可谓是量身定制。**RAG**利用广泛的数据库和大语言模型(LLM)的动态功能可生成富有见解、准确的结果。



RAG的最佳用例

RAG架构的多功能性极强，适用于从对话式AI，内容创建，到搜索引擎中的复杂查询解决等各种应用场景。其独特之处在于能够根据检索到的实际数据生成回答，因此特别适合以下场景：

客服系统：

为客户查询并提供准确且与上下文相关的回答。

研究分析工具：

提炼复杂信息，形成全面的总结。

教育平台：

根据广泛的知识库提供详细的解释或回答。

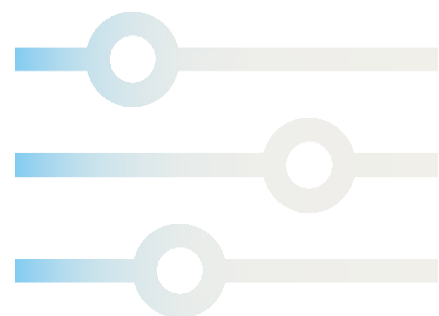
内容生成：

跨不同领域生成丰富、信息量大且相关的内容。

RAG的架构

RAG架构主要由两大部分组成：数据准备和RAG系统本身。数据准备涉及连接到原始数据源，使用上下文信息标注数据，将文档分为更小的连贯单元，以及将这些单元转换为向量嵌入。分单元可以提高连贯性，而嵌入可以使RAG生成器实现更好的语义匹配和相关知识的检索。

RAG系统接受用户提示，在嵌入式系统中搜索相关段落，然后将它们发送到LLM以生成回答。人类的参与在数据准备过程中及在RAG系统中都至关重要：在数据准备中，需要人类将领域专业知识和上下文信息添加到原始数据中；在RAG系统中，需要人类增强向量检索相关性并提供提示或回答以保证产出质量。



在RAG中创造卓越：关键决策

RAG的有效性取决于[其架构中各个方面的细致决策](#)：

数据准备：

选择数据并将其分为连贯的单元，同时不丢失实质上下文信息。澳鹏在数据标注和分块方面拥有丰富的专业知识，可确保数据上下文信息完整无缺且标记准确。

相关性和排名：

用于确定检索到的数据与用户提示的相关性及其后续排名的算法。澳鹏的人机协同方法可增强检索到的数据的相关性和排名。

嵌入技术：

选择正确的方法将文本分块转换为嵌入，以实现高效检索。

确保RAG输出的质量

为了保持RAG输出的高质量和准确性，需要考虑以下几个方面：

全面的数据标注：

需要人类准确地标注和分类数据。

持续评估：

定期根据各种提示测试模型以确保一致性和可靠性。

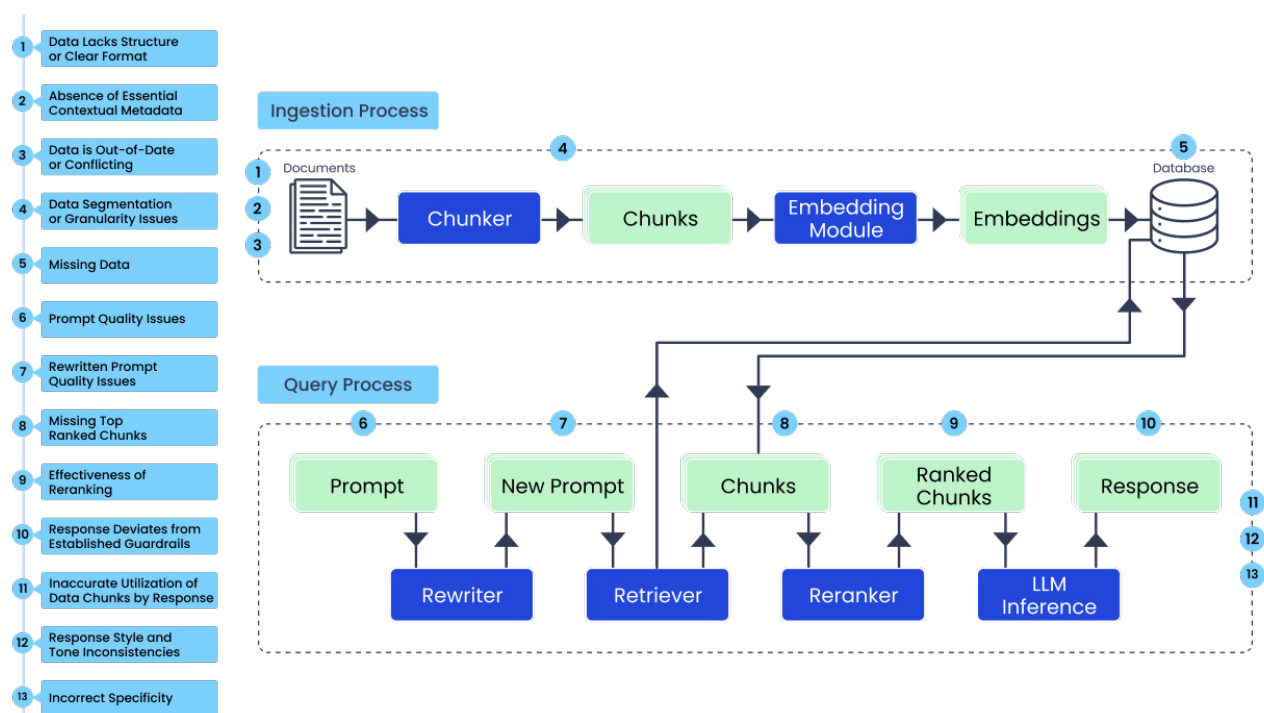
减少偏见：

主动寻找并消除数据集和模型回答中的偏见。



RAG优化的数据准备

实施RAG需要细致地准备数据源，因为模型将从数据源中学习和检索上下文。由于数据源多种多样，往往同时包含结构化数据和非结构化数据，因此，对大型知识库进行清理、构建和优化并导入向量数据库可能会面临诸多挑战。



导入过程

1. 数据源:

根据主题专家(Subject Matter Expert)的意见, 选择企业结构化和非结构化数据存储库中符合用例需求的相关子集。

2. 数据清理:

数据输入向量存储库前, 严格进行数据过滤和去重处理。

3. 隐私/PII:

确保对PII元素进行适当处理, 包括检测、过滤、编辑以及适当替换为合成数据。

4. 文本提取:

从非结构化和半结构化数据源中提取干净、可用的文本。

5. 文本规范化:

应用规范化规则, 标准化拼写、语法、测量和一般命名法。

6. 分块策略:

将文档拆分为可检索单元, 围绕关键信息保留完整、相关的上下文。

7. 实体识别和标注: 为分块添加丰富的元数据, 如数据上的源详细信息、主题和关键实体。

查询过程

在RAG中, 查询过程包含多个关键步骤, 以确保回答的准确性和相关性:

段落排名:

需要人类准确地标注和分类数据。

提示工程与设计:

构建提示, 以有效获取和利用检索到的上下文, 同时与输出回答的所需风格和语调保持一致。

持续评估和优化

为有效识别和监控性能差距, 持续的测试、评估和优化尤为重要。我们可以采用组件级评估来精准定位具体问题, 同时, 利用端到端评估, 基于目标用例来全面考量整个系统的性能表现。

RAG系统中的人类专业知识

在实施检索增强生成(RAG)系统时，开发人员经常会面临数据质量低和数据重复等难题。这些难题可能会影响系统提供准确、相关回答的能力。人类专业知识在解决这些问题方面发挥着至关重要的作用。通过利用人类的知识和判断，开发人员可以改进RAG系统的性能，确保它能够有效地处理复杂的现实世界需求。

为了保持RAG系统能够提供精准解决用户查询的高质量回答，对RAG系统进行专家管理和持续更新必不可少。这也进一步凸显人类干预在RAG系统开发过程中的核心价值。

人类如何改进RAG结果

在改进RAG结果的过程中，人类发挥着举足轻重的作用。他们不仅确保了数据格式规范，从而实现有意义的分割，还巧妙地融入上下文元数据，极大地增强了数据分块与检索的相关性。为了确保向量存储内容的准确性和时效性，他们会执行严格的质量保证流程。同时，他们还会对分块方法进行持续优化，确保能够精准捕捉语义的上下文。

专家会定期审核和更新系统，使用回退机制解决缺少数据问题。他们创建反映真实用户行为的高质量提示，修改提示以保持用户意图，从而确保系统的有效性。通过性能评估和调整，人类提高了算法提取最相关数据块的能力。

由人类定期进行问答评估可防止偏离指南并确保遵守回答的规范。他们改进检索算法并提供有针对性的训练数据以提高数据块利用率。通过编辑训练数据集，人类可以确保回答风格和语调的一致性，同时调整系统参数并添加注释以增强回答的具体性。

人类专业知识在RAG系统中的作用

在整个RAG架构中，人类监督对于以下几方面至关重要：

数据标注和分块：

确保上下文信息完整无损且标注准确。

确保上下文完整性：

解读超出AI能力的细微差别和上下文。

模型训练和调整：

通过细致的理解和反馈完善模型。

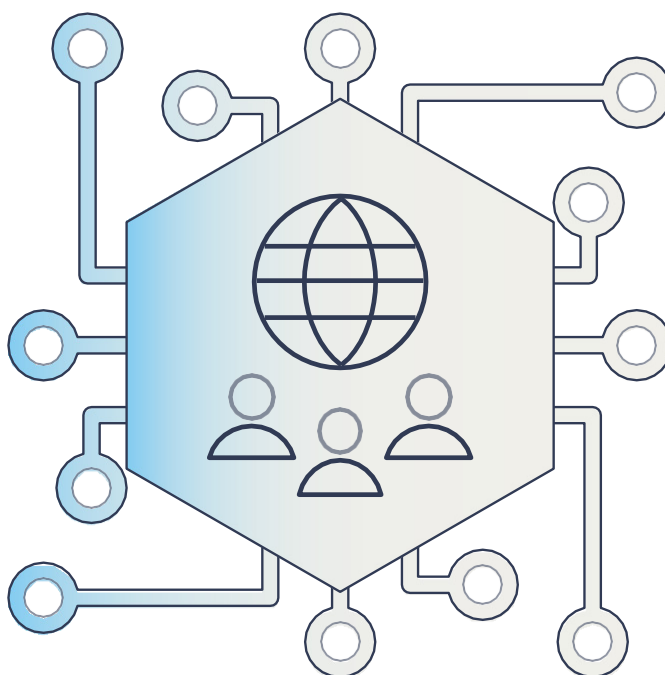
防止错误：

持续监督，实时识别和纠正错误。

质量控制：

监督模型输出的准确性、有用性和安全性。

通过将自动评估指标与人类判断相结合，RAG系统实现了高准确性、相关性和可靠性，这对于现实世界中的应用场景至关重要。



自动评估指标：监控RAG的准确性和相关性

虽然人类的参与在提高RAG系统的质量和完整性方面不可或缺，但自动评估指标在持续监控这些架构的性能方面同样具有举足轻重的地位。通过这些指标，AI团队能够快速确定需要改进的领域，确保RAG在实时应用中始终保持卓越的性能与效率。RAG流程中使用的核心自动评估指标涵盖检索阶段和生成阶段。

检索指标

上下文回忆：

衡量检索组件成功返回的相关分块的百分比。

上下文准确性：

评估检索到的分块被归类为相关分块的准确性。

生成指标

生成阶段指标关注输出的真实性和与提示的相关性，确保生成的文本符合事实正确性和相关性。

真实性：

衡量所生成的句子/短语与检索到的相关分块的实际正确性不符的数量。

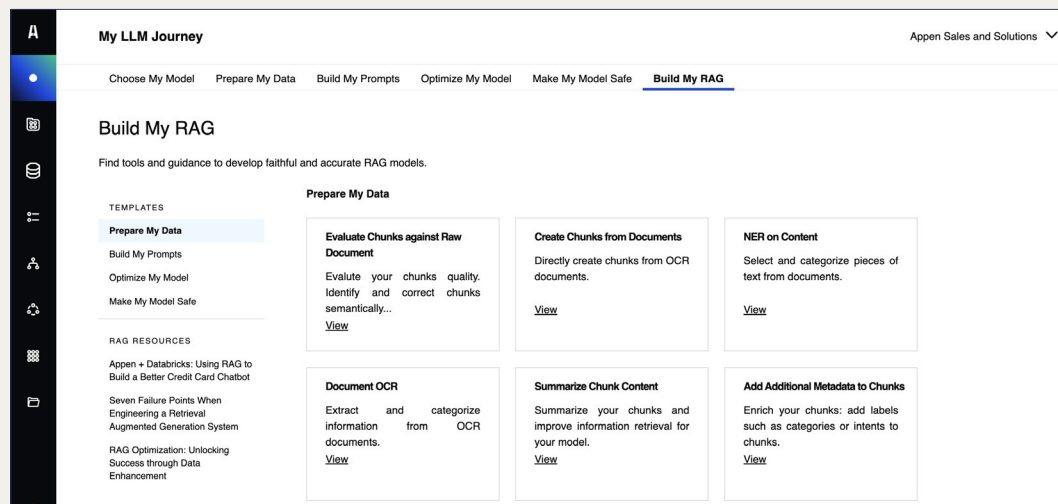
相关性：

评估所生成的句子/短语与用户查询无关的数量，确保系统的回复重点突出、主题鲜明。

澳鹏的“Build My RAG” （构建我的RAG）功能

为了更好地支持团队基于专有数据构建AI应用，我们在[AI数据标注平台 \(ADAP\)](#)中发布了“[构建我的RAG](#)”功能。这一新功能可帮助团队创建高质量的RAG系统。

[构建我的RAG](#)功能提供了一套全面的模板，涵盖了基本任务，例如从复杂的PDF中删除重复数据或提取数据以确保向量数据库导入的可扩展性，使用来自其他系统的标注，或标注丰富源数据以提高检索质量，以及整理黄金数据集以帮助评估端到端系统。[观看我们的演示视频](#)。



主要步骤

1. 准备我的数据:

利用模板来分割和丰富您的文档，确保您嵌入的数据连贯且相关。

2. 构建我的提示prompt:

使用专用模板设计有效的提示。这些模板指导您设计问题和命令并帮助评估其质量。

3. 优化我的模型:

评估、排序并优化您的RAG模型的回答，通过识别和纠正差异来提高准确性。

4. 确保我的模型安全:

通过严格的红队测试流程确保稳健性和可靠性。使用AI聊天反馈来测试性能并识别潜在漏洞。

模板提高效率

我们预先构建的模板可帮助您创建高效、可靠的RAG模型，以满足您的特定需求。它们有助于：

- 根据源文档评估分块以确保完备性、相关性和完整性。
- 通过总结分块实现有效的信息检索，并为它们添加丰富的类别或意图等元数据。
- 从OCR文档中选择和分类信息，使用NER标记和分类文本，并提取必要详细信息。
- 将元数据分配给原始文档，改进分块文本以确保准确性，并对相似的分块进行重复数据删除以保留多样化和独特的信息。

结语

有效的AI系统需要在整个AI生命周期中进行人类的交互。**RAG**架构通过整合特定领域的数据增强基础模型的功能。这种方法依赖于领域专家实施监督，确保AI的诚信、实用性和安全性。

澳鹏的AI数据平台通过在数据准备、模型评估和优化方面提供至关重要的人类专业知识和专家监督，助力企业部署**RAG**模型。企业可以使用自己的专家、澳鹏的全球众包资源或两者结合使用。

我们的平台涵盖数据采集、清理、标注和优化，可解决数据不一致和信息过时等问题。凭借澳鹏在数据标注和质量保证方面的经验，可以帮助您充分发挥**RAG**系统的潜力，确保准确性、相关性和人类的洞察力。

利用澳鹏的专业知识

联系welcome@appen.com，了解我们的专业知识和先进平台如何帮助您加速**RAG**之旅。