

的落地难度。

除了基础实现外，我们还可以针对垂直领域进行深度优化，如为法律合同解析设计专用提示模板，或为医疗文献问答配置专项模型参数，进一步提升系统在特定场景下的表现。这些优化与扩展，将为 RAG 的企业应用带来更多可能。

3.4.4 小结

RAG 技术的核心优势在于它能够利用海量外部知识，生成更加精准、丰富的内容。而要充分发挥 RAG 的威力，前提是高效地组织和检索这些知识。文本分割作为 RAG 数据使用的首要环节，需要在语义保留和计算效率之间取得平衡。我们首先介绍了多种分割策略，包括基于长度的文本分割、基于语义的文本分割等，并分析了它们的适用场景和优缺点。

同时，分割后的数据组织形式也影响着 RAG 系统的性能。通过构建合适的索引结构、选用高效的数据组织结构和查询优化技术，我们可以最大限度地提升数据的可访问性和检索效率。LlamaIndex、LangChain 等先进工具的出现，进一步降低了 RAG 的应用门槛。

3.5 总结

本章全面讨论了 RAG 系统中的两大关键问题：向量数据库与数据处理。向量数据库作为支撑 RAG 实现高效知识检索的核心组件，其选型和应用是 RAG 落地的首要问题。我们系统阐述了向量数据库的基本概念、工作原理、技术特点，并对比了主流的向量数据库产品。同时，我们提供了如何权衡数据规模、召回率、计算资源等因素并选择适合的解决方案的参考建议。