

Topic 11: RAG

1. Hugging Face Tutorial: Code a simple RAG from scratch:
<https://huggingface.co/blog/ngxson/make-your-own-rag>
2. Chunking in RAG applications:
<https://stackoverflow.blog/2024/12/27/breaking-up-is-hard-to-do-chunking-in-rag-applications/>
3. Weaviate Blog: Chunking Strategies to Improve Your RAG Performance:
<https://weaviate.io/blog/chunking-strategies-for-rag>
4. NVIDIA Technical Blog: Finding the Best Chunking Strategy for Accurate AI Responses:
<https://developer.nvidia.com/blog/finding-the-best-chunking-strategy-for-accurate-ai-responses/>
5. Milvus (Zilliz) Blog on How to Choose the Right Embedding Model for RAG:
<https://milvus.io/blog/how-to-choose-the-right-embedding-model-for-rag.md>
6. Elastic Blog: RAG Explained:
<https://www.elastic.co/blog/retrieval-augmented-generation-explained>
7. RAG Evaluation in Practice
<https://kinde.com/learn/ai-for-software-engineering/best-practice/rag-evaluation-in-practice-faithfulness-context-recall-answer-relevancy/>
8. What is RAG? <https://www.youtube.com/watch?v=T-D1OfcDW1M>