

# Hands-On: Build a Simple RAG System

## Instructor

Dipanjan Sarkar

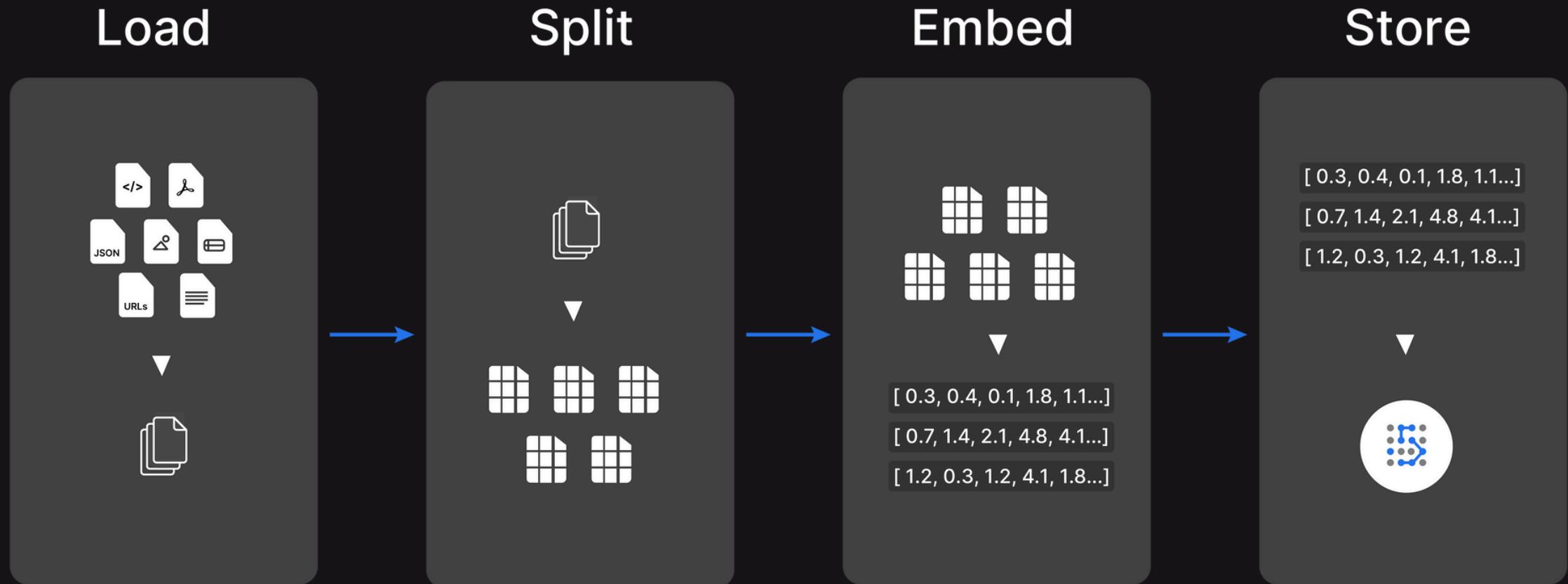
Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

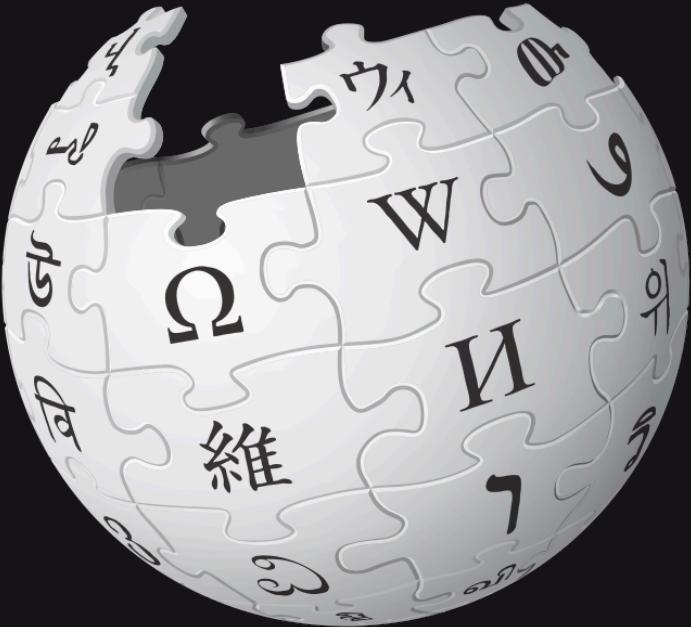
Published Author



# RAG Workflow - Step 1 - Data processing and Indexing



# Data Source



Text Article JSON

**Deep Residual Learning for Image Recognition**

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun  
Microsoft Research  
[kaiming.xiangyu.zhang@intel.com](mailto:kaiming.xiangyu.zhang@intel.com)

**Attention Is All You Need**

---

**TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE**

Alexey Dosovitskiy<sup>\*†</sup>, Lucas Beyer<sup>\*</sup>, Alexander Kolesnikov<sup>\*</sup>, Dirk Weissenborn<sup>\*</sup>, Xiaohua Zhai<sup>\*</sup>, Thomas Unterthiner<sup>\*</sup>, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby<sup>\*‡</sup>  
<sup>\*</sup>equal technical contribution, <sup>†</sup>equal advising  
Google Research, Brain Team  
[{adosovitskiy, neilhoulsby}@google.com](mailto:{adosovitskiy, neilhoulsby}@google.com)

---

**ABSTRACT**

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure attention approach can be competitive. Specifically, we propose to perform attention on multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTFB, etc.). Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.<sup>1</sup>

**1 INTRODUCTION**

Deep convolutional neural networks have become the de-facto standard for image classification. A series of breakthroughs [50, 40] have led to networks that are deeper than ever before. However, the depth of representations in these networks is still limited by the number of layers. The recent success of ViT [41, 44] reveals that not only the leading results in image classification [17, 18, 19] with a depth of sixteen layers, but also the best models in visual recognition [17] with a depth of sixteen layers, are achieved by using attention mechanisms.

The dot product attention mechanism is based on a series of dot products between query, key, and value vectors. It is simple, yet effective, and has been successfully applied to many other tasks such as machine translation [41], document summarization [17], and question answering [18].

Self-attention mechanisms, in particular Transformers [Vaswani et al., 2017], have become the method of choice in natural language processing (NLP). They were originally designed for large text corpora and then fine-tuned on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers’ computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

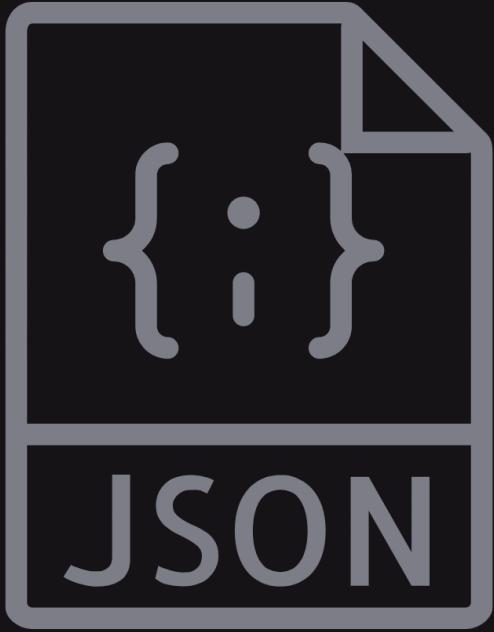
In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNNs with self-attention (Wu et al., 2019; Xie et al., 2019), or replacing the convolution entirely (Ranjan et al., 2019; Wang et al., 2020). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state-of-the-art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image

<sup>1</sup>Equal contribution. This work was partially funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825308 (Project DECA). The authors would like to thank the anonymous reviewers for their useful comments and suggestions.

Research Paper PDFs

# Data Loader

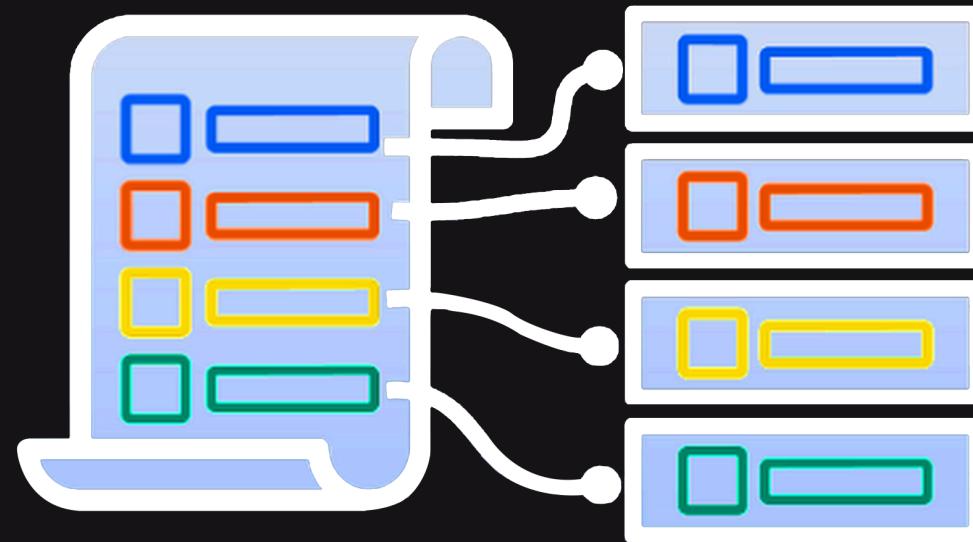


JSON Loader



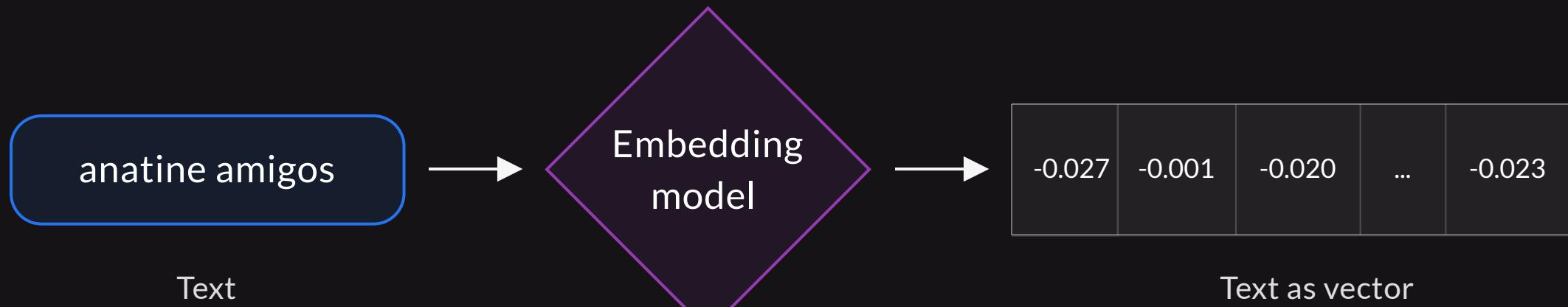
PDF Loader

# Chunking Strategies



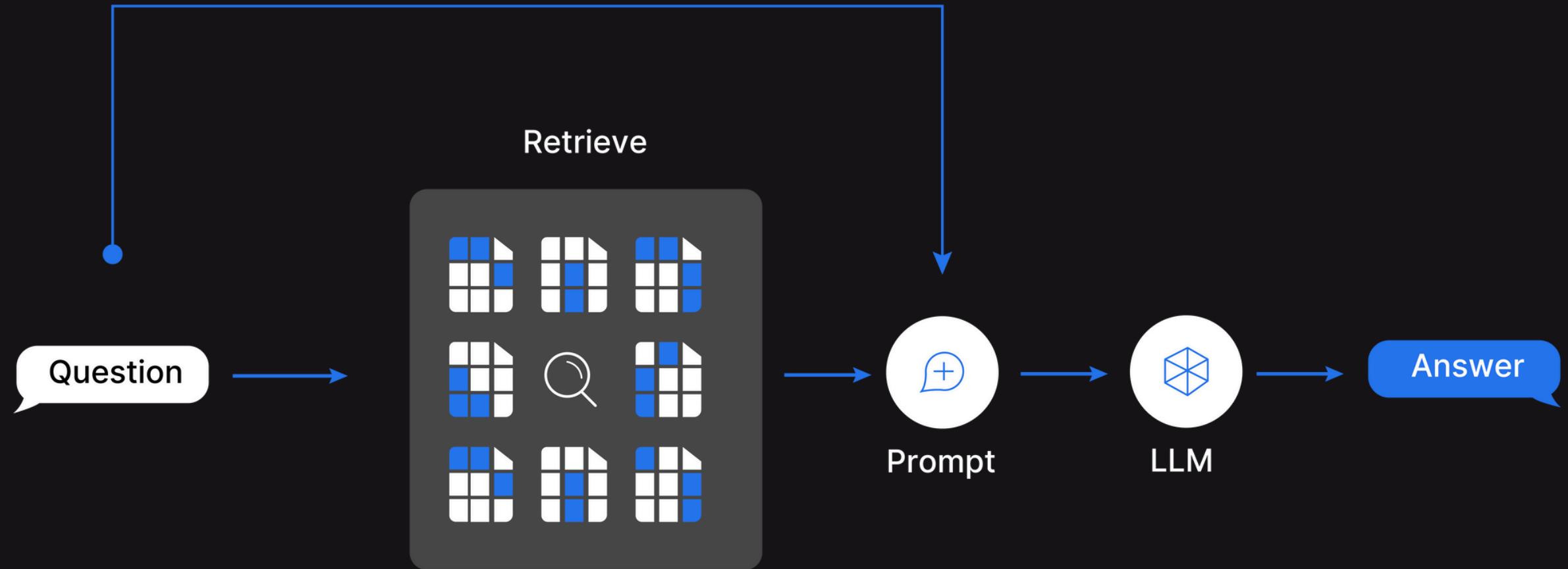
Recursive Character Text Splitting & Chunking

# Embedder Model

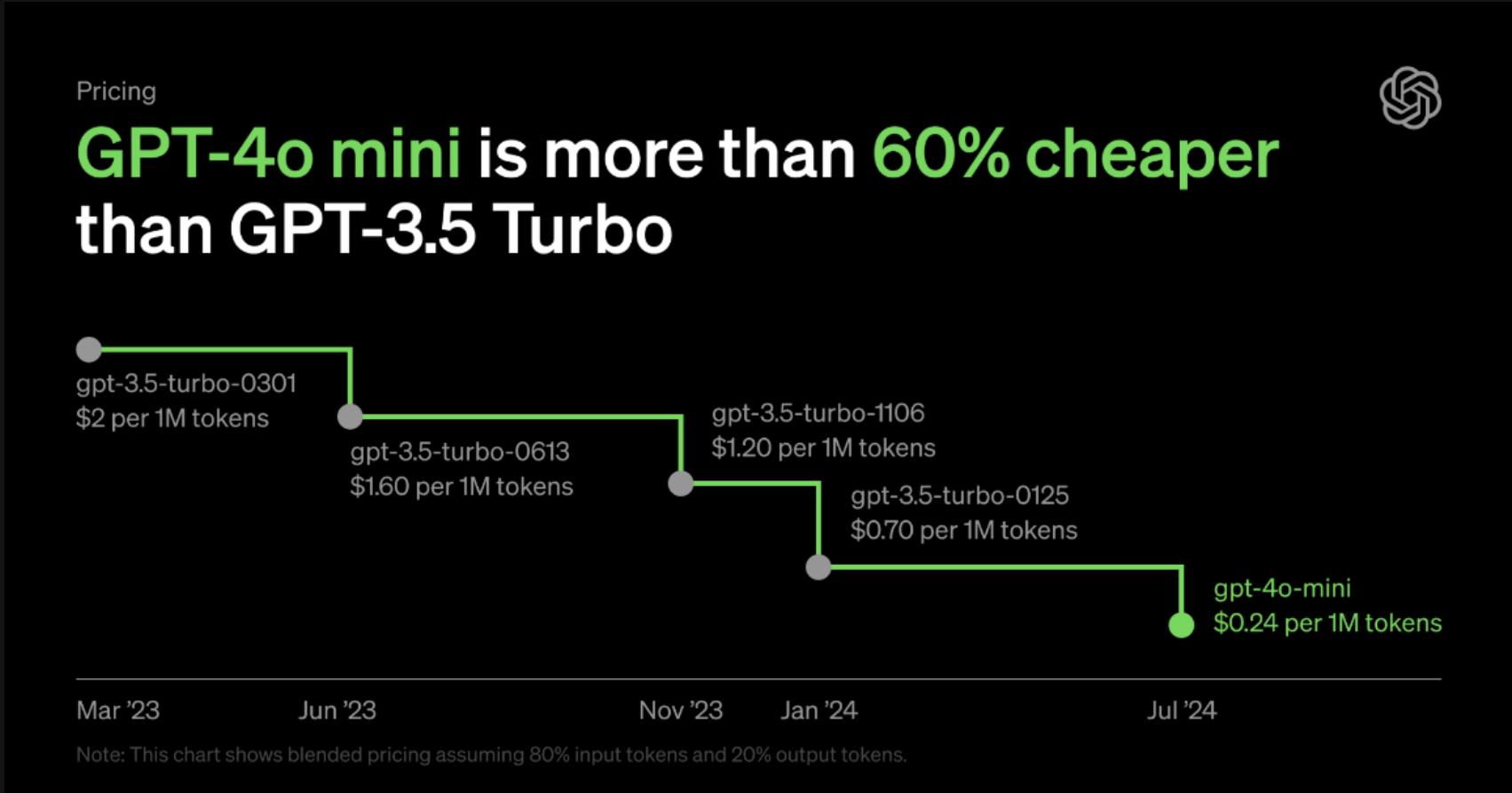


OpenAI Embedder

# RAG Workflow - Step 2 - Retrieval and Response Generation



# LLM for Response Generation



# Thank You

---