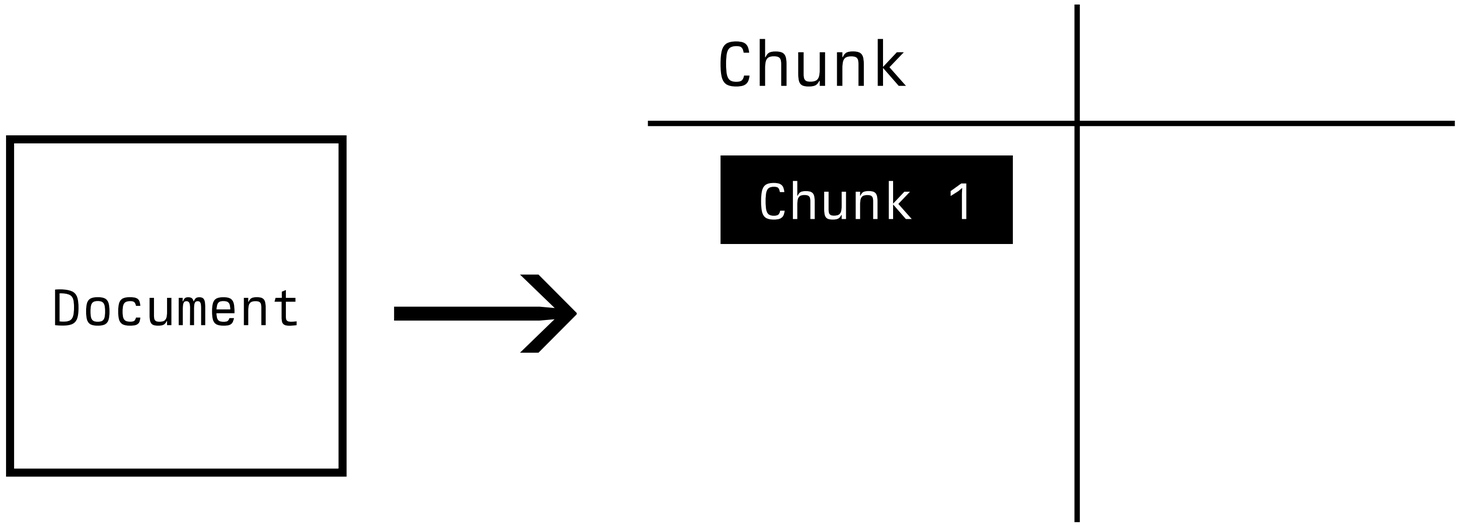
FlashRAG

Blazing-fast RAG inference for the impatient

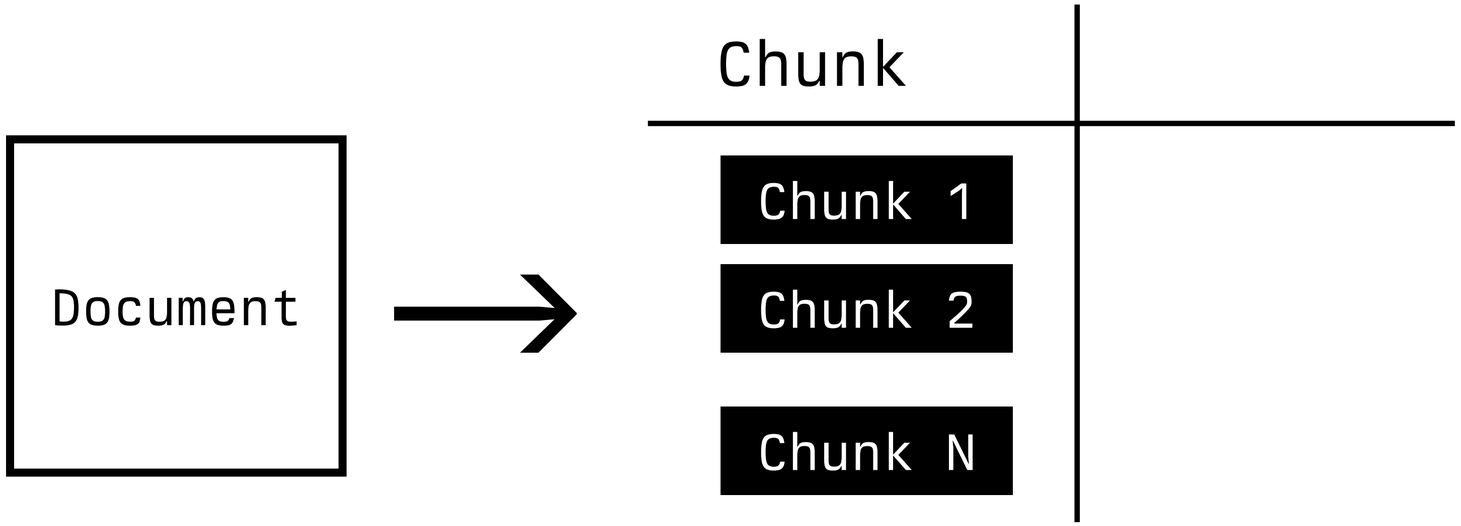
Vitaly Kleban, vk@cyber.fund, @vkleban

Naïve RAG Inference

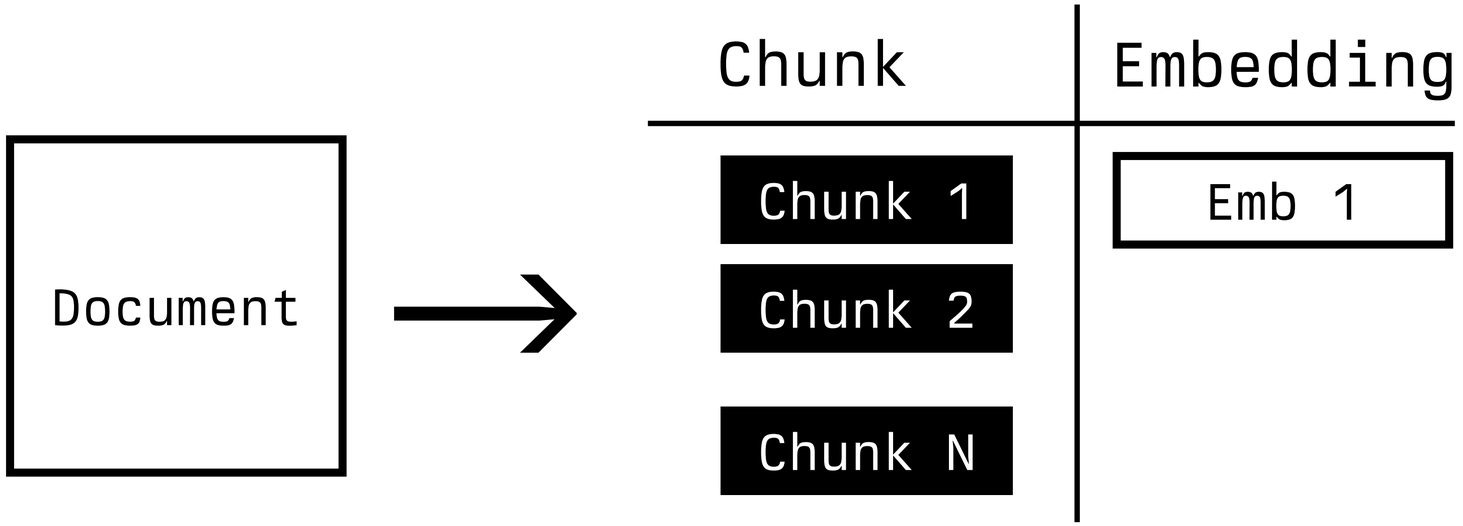
Naïve RAG Inference



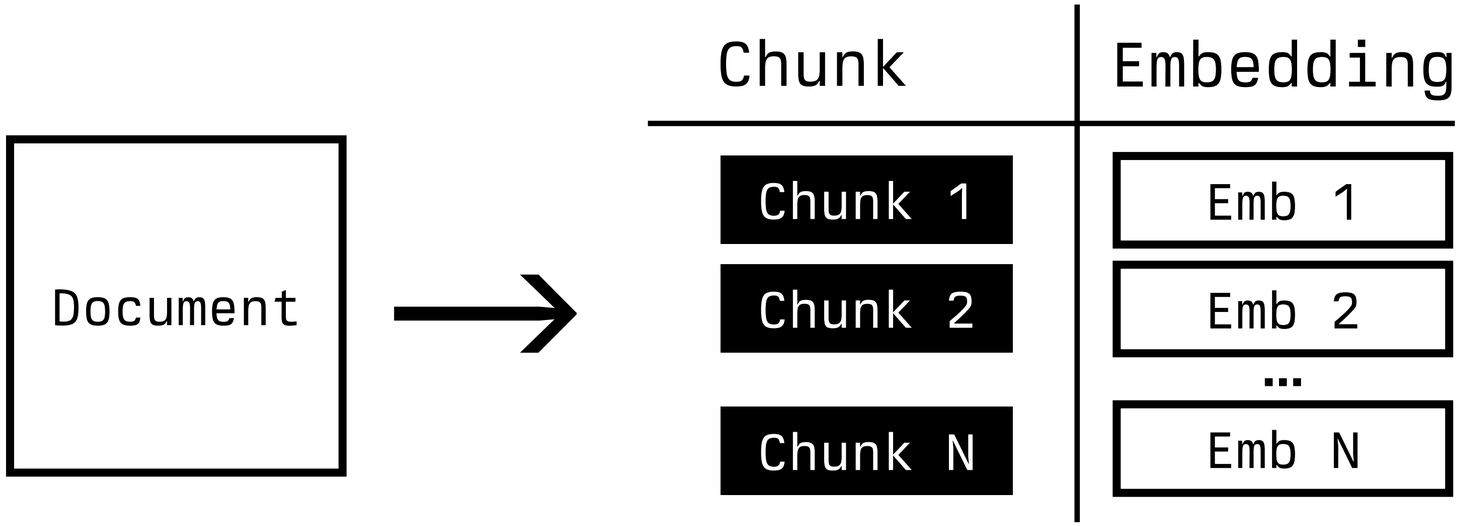
Naïve RAG Inference



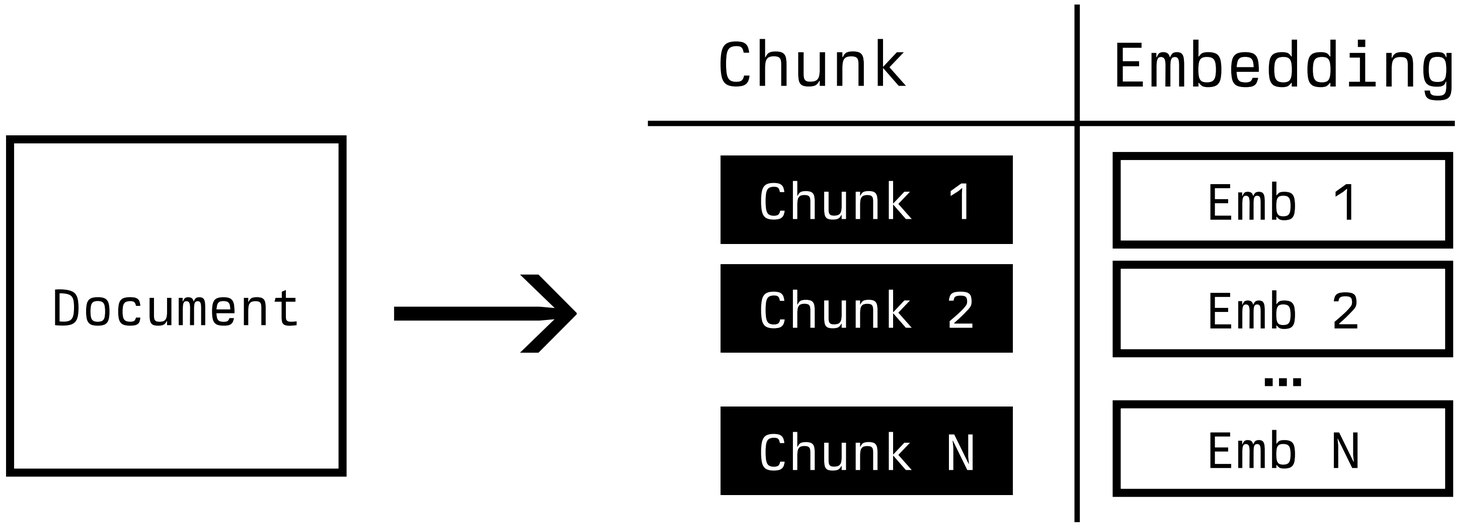
Naïve RAG Inference



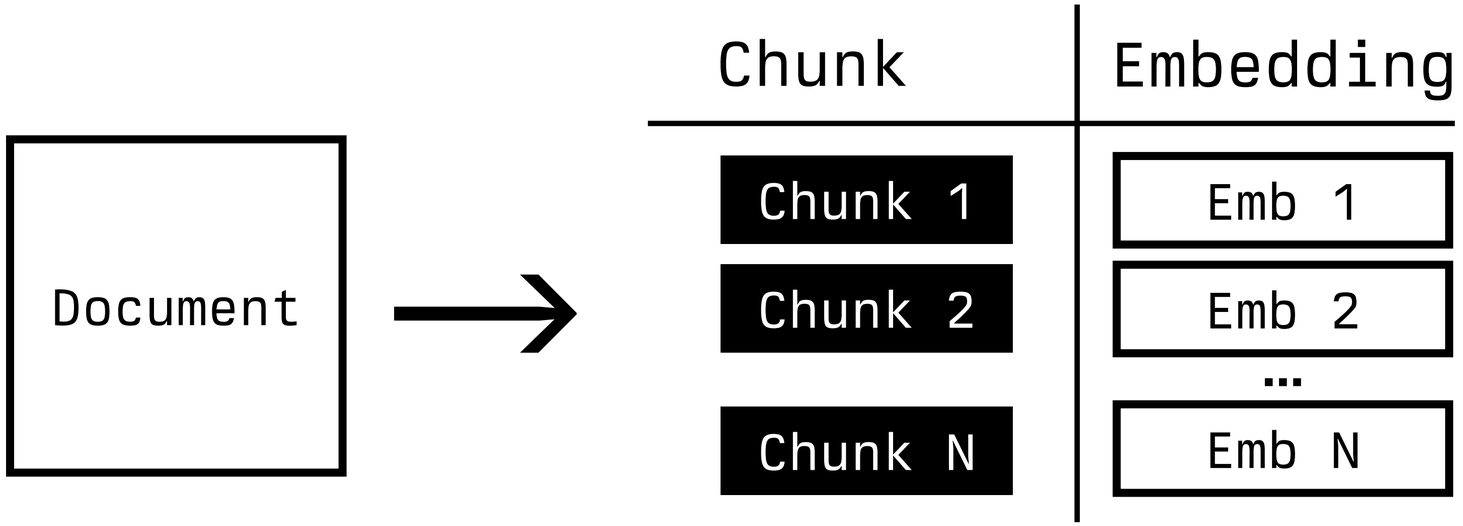
Naïve RAG Inference



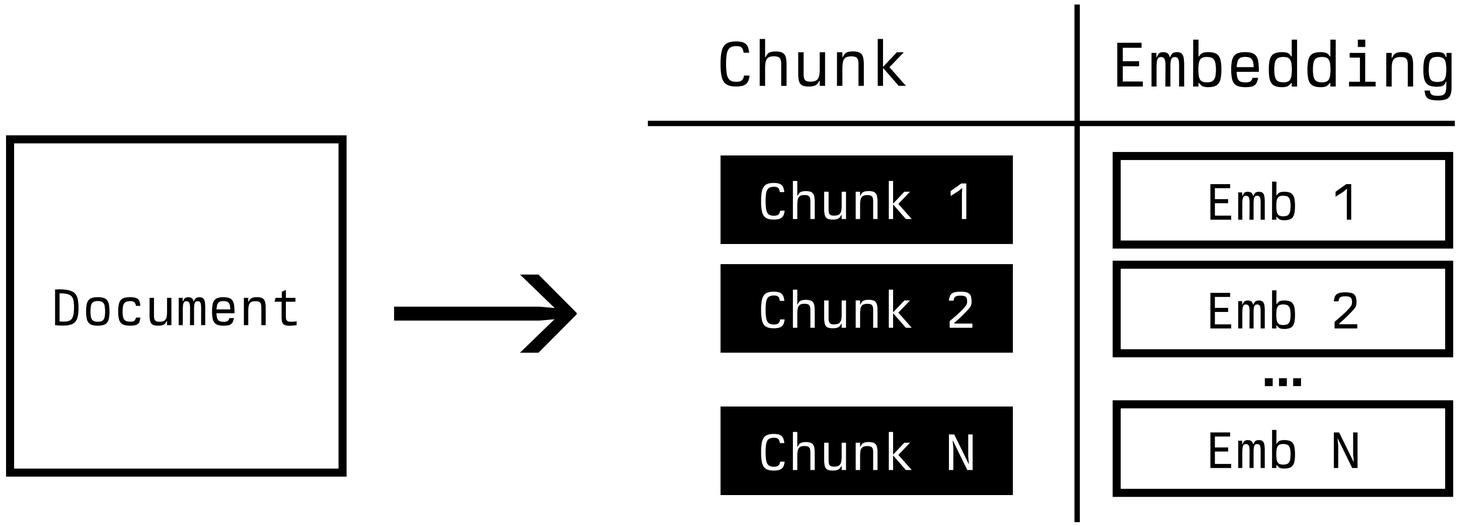
Naïve RAG Inference



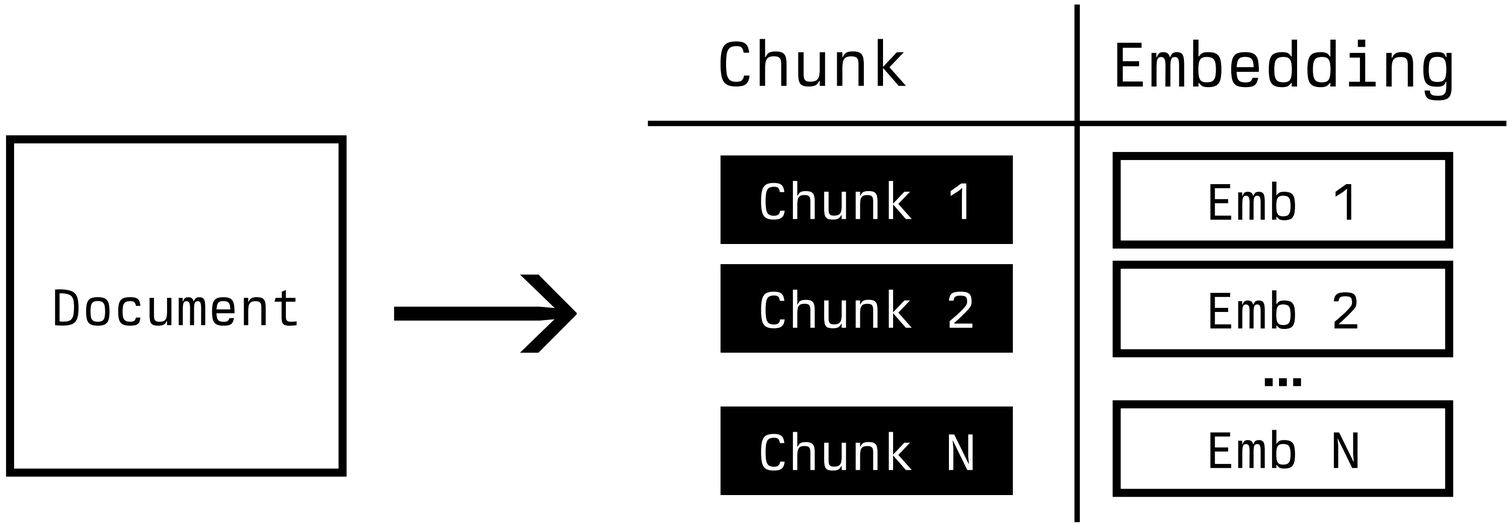
Naïve RAG Inference



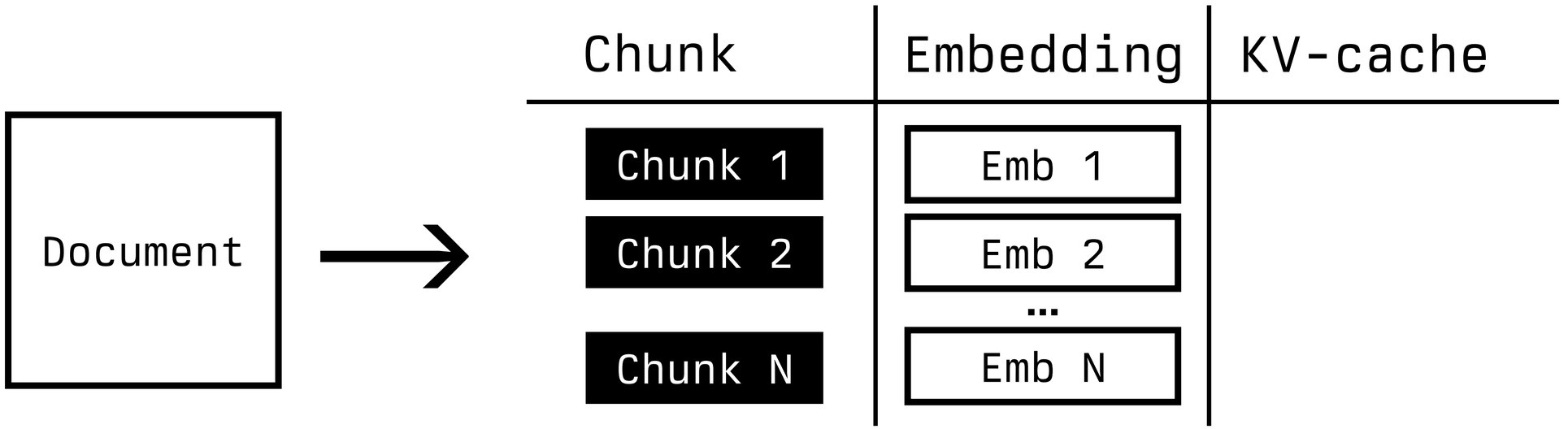
Naïve RAG Inference



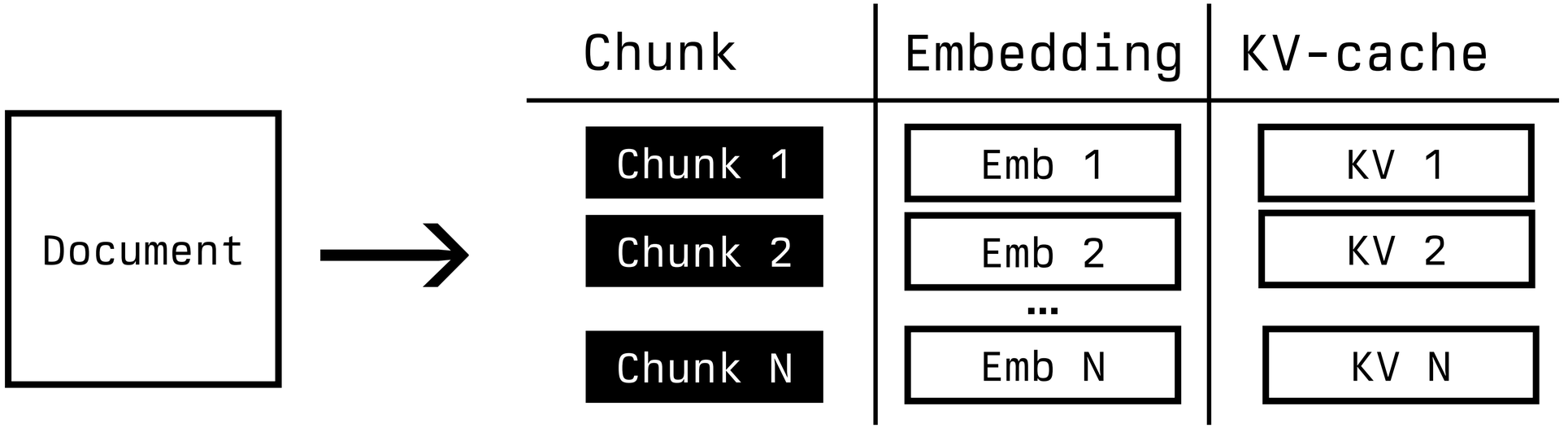
Flash RAG Inference



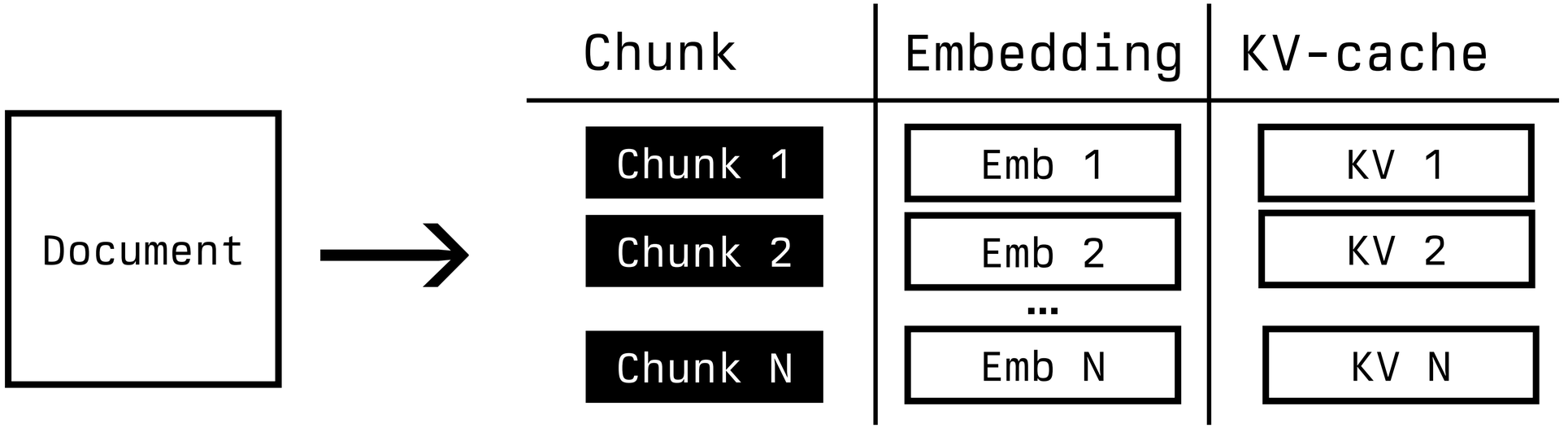
Flash RAG Inference



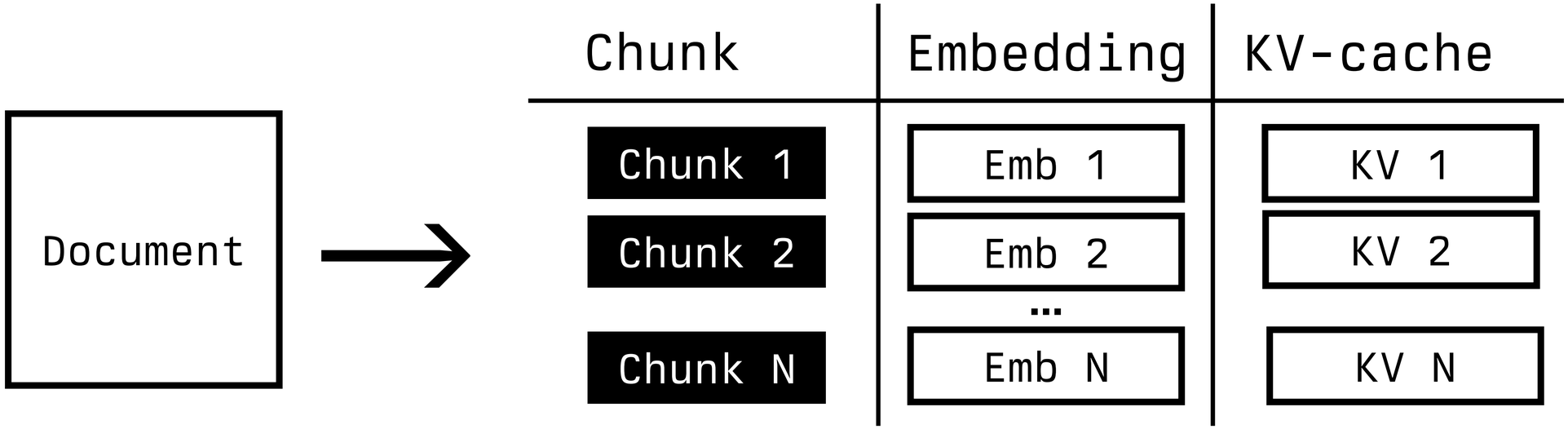
Flash RAG Inference



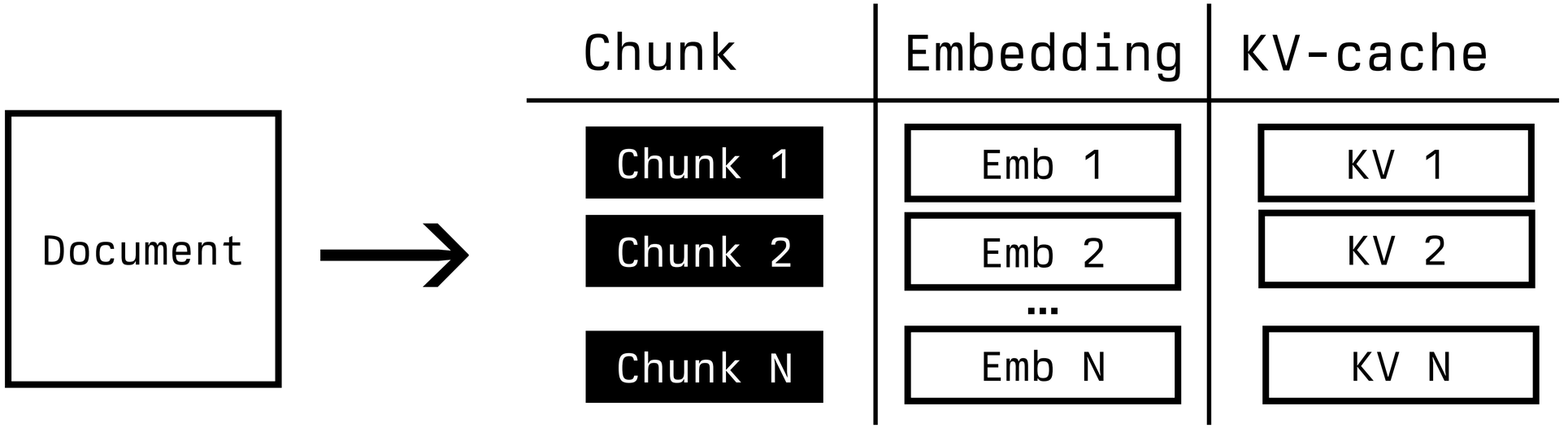
Flash RAG Inference

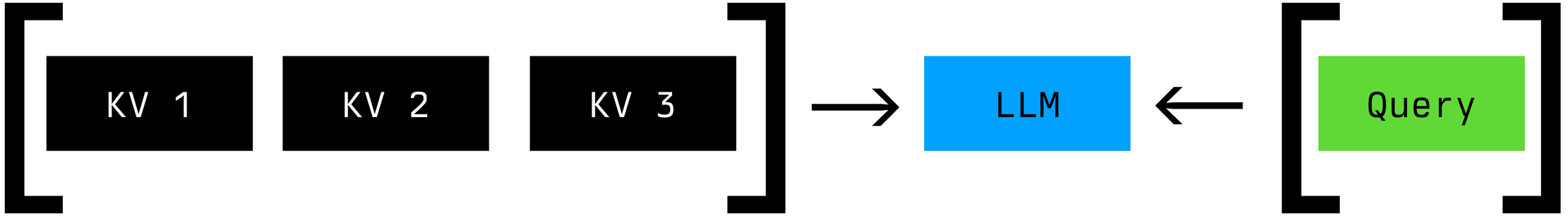


Flash RAG Inference

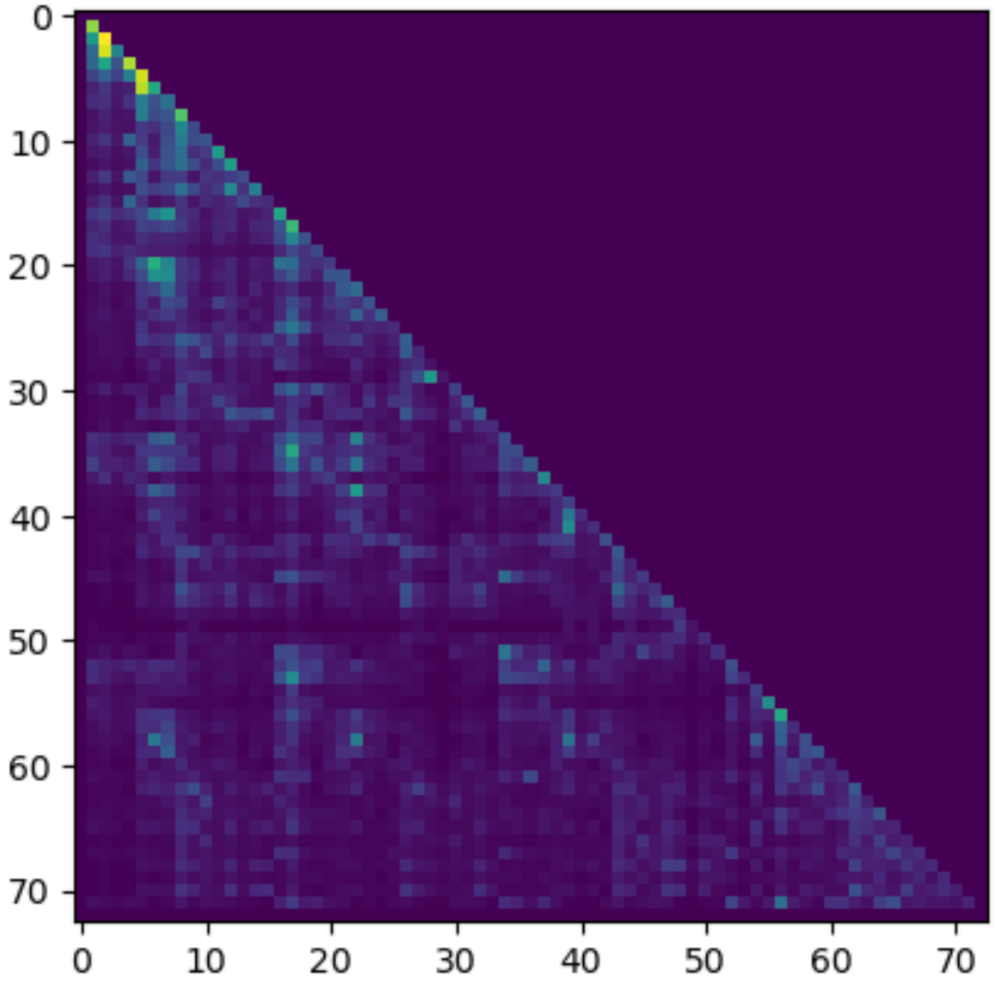


Flash RAG Inference





Why is it possible?



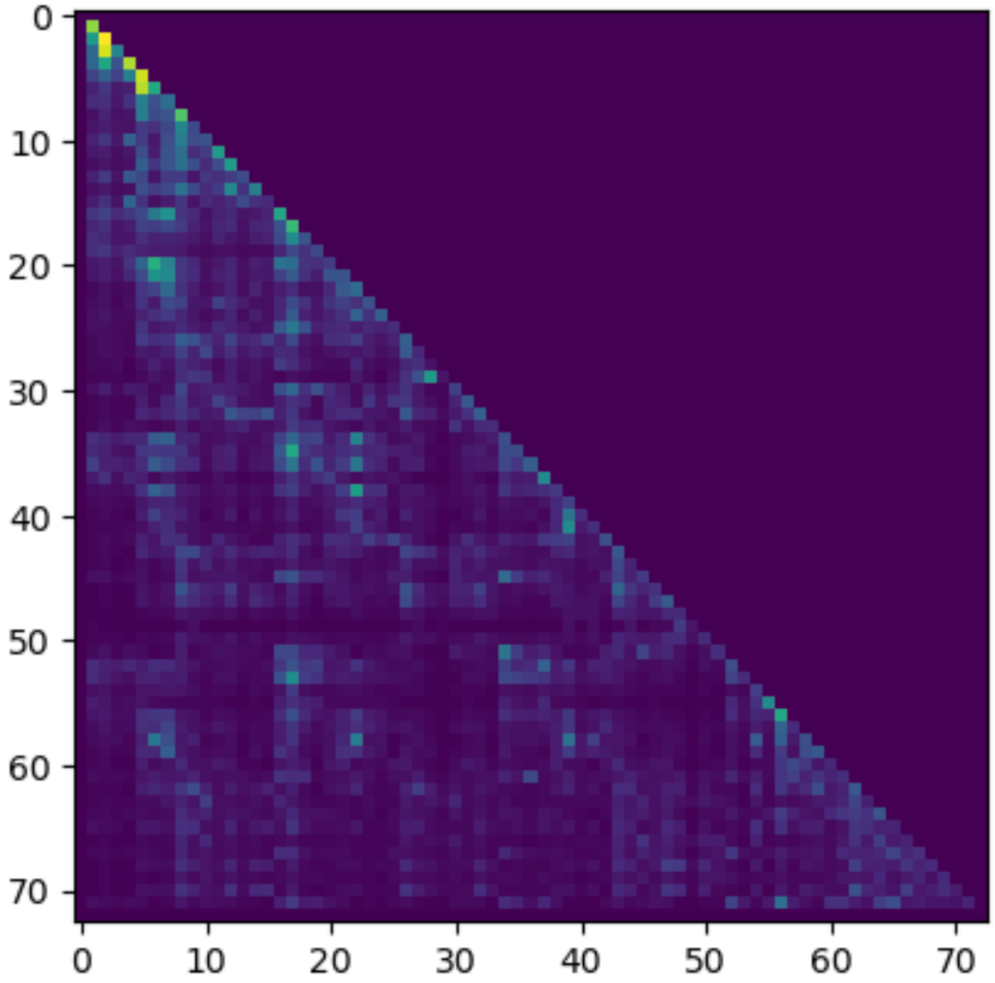
Chunk 1

Chunk 2

Chunk 3

Query

Why is it possible?

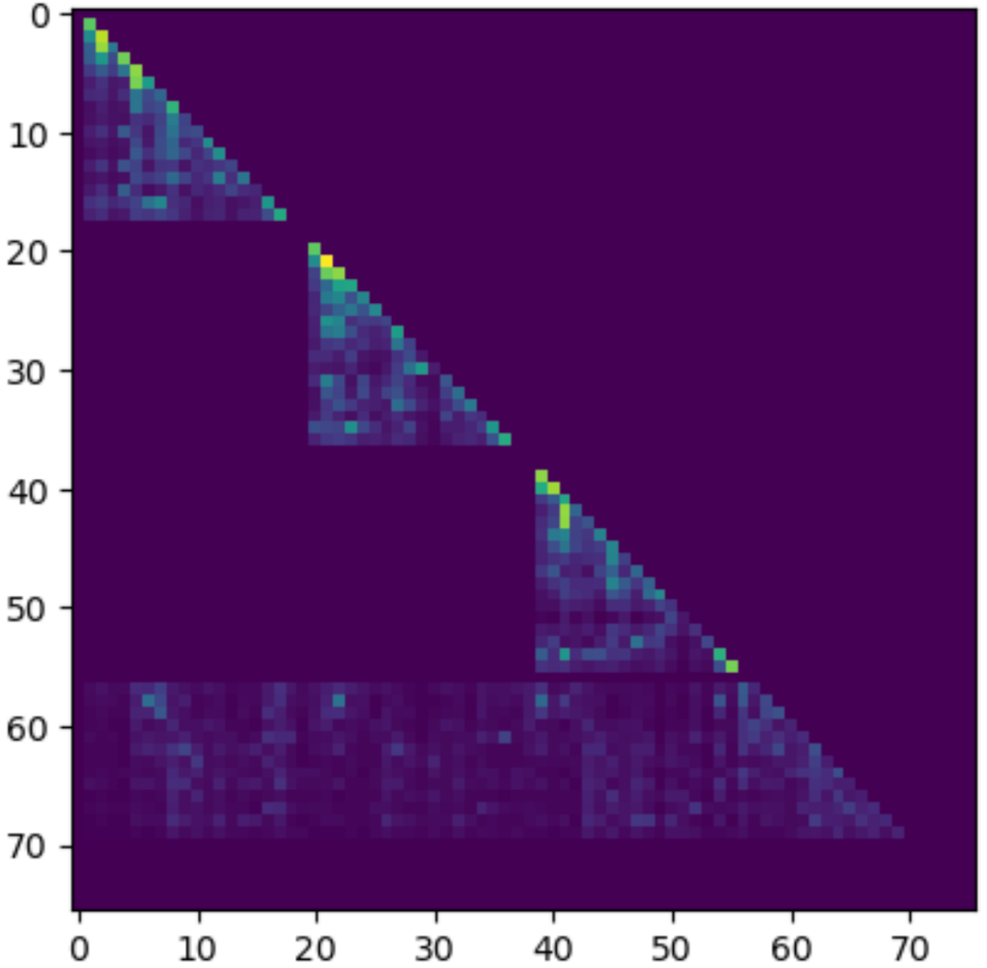


Chunk 1

Chunk 2

Chunk 3

Query



Why is it possible?