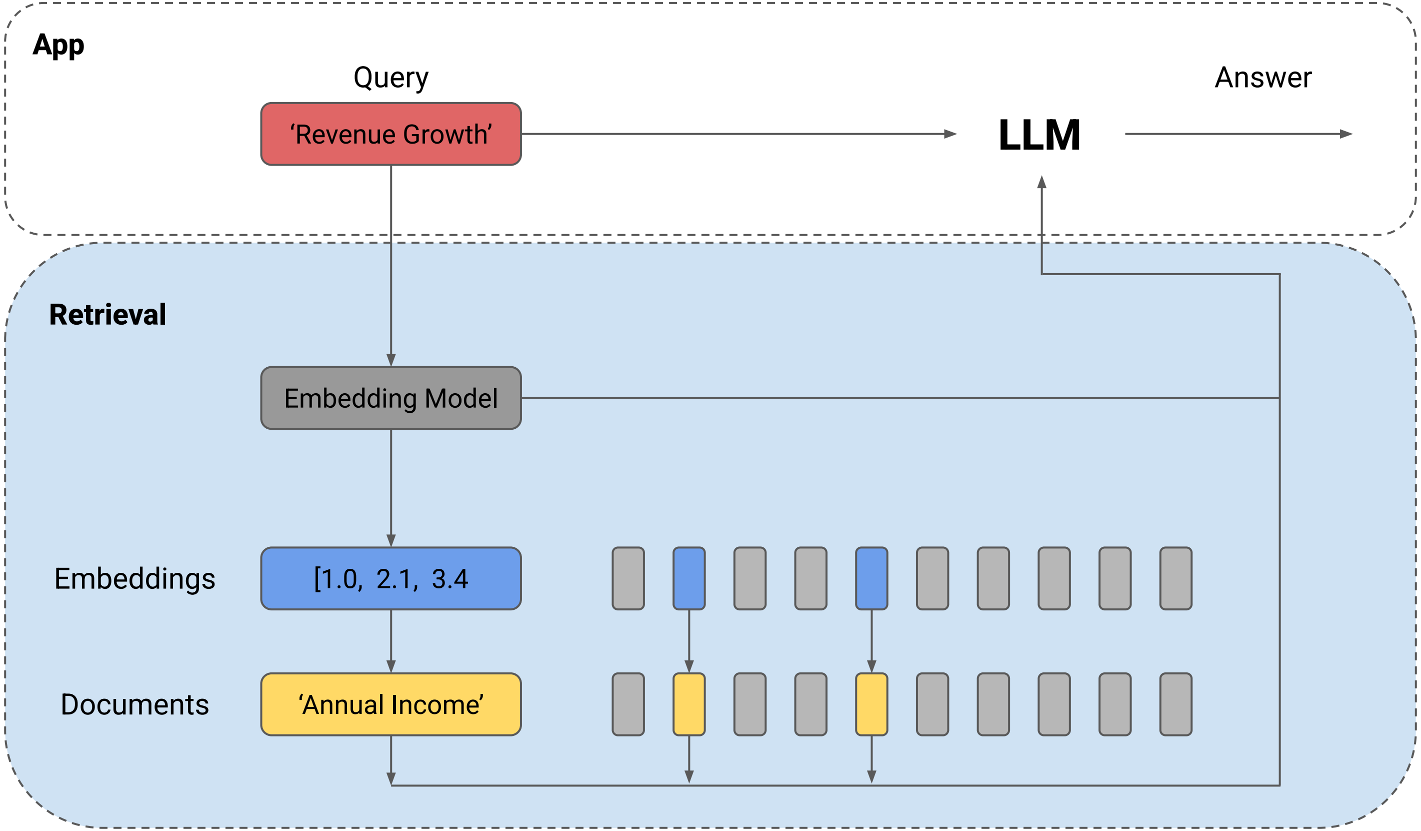
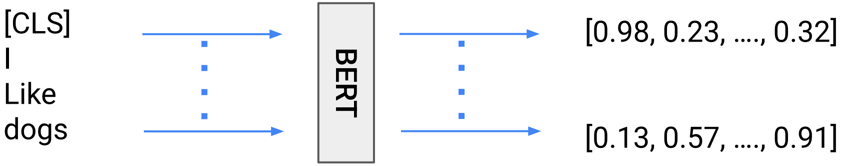
Advanced Retrieval for AI with Chroma

Retrieval Augmented Generation



Sentence Transformer

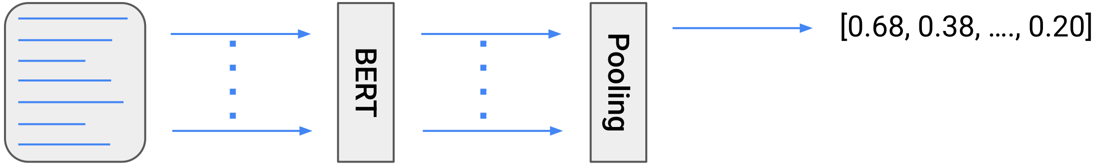


[CLS]
I
Like
dogs

BERT

[0.98, 0.23, …., 0.32]

[0.13, 0.57, …., 0.91]

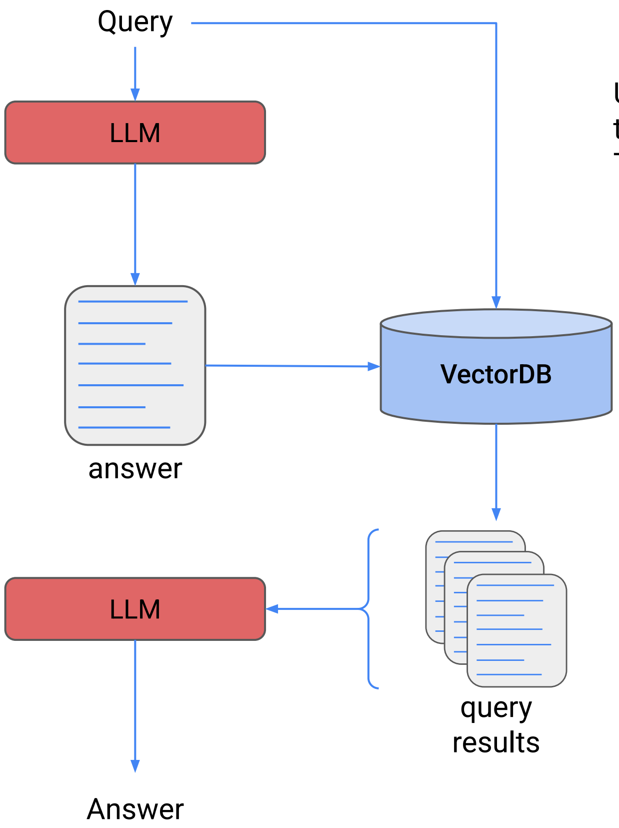


BERT

Pooling

[0.68, 0.38, …., 0.20]

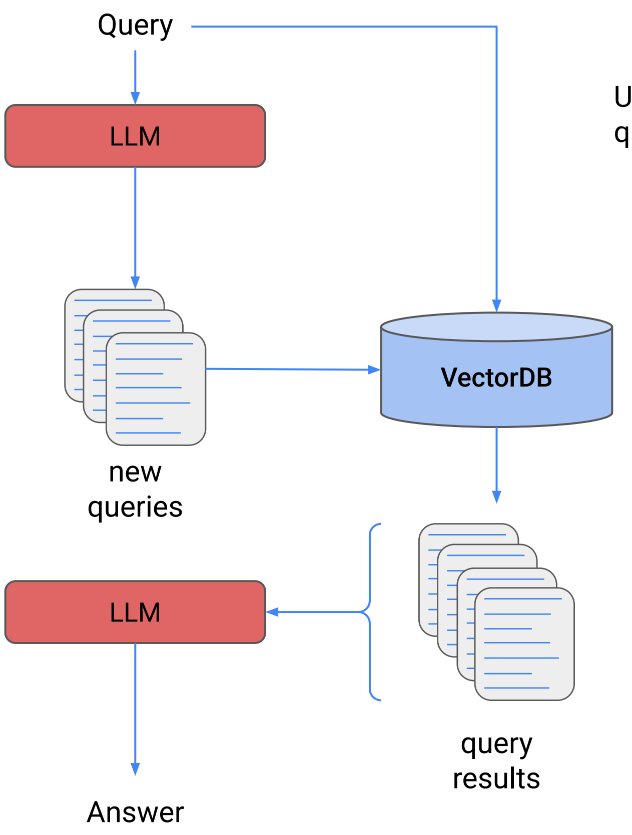
Expansion with Generated Answers



Use an LLM to suggest an answer
to the question.
This provides relevant context

Retrieve n results for original query
with additional context

Expansion with Multiple Queries

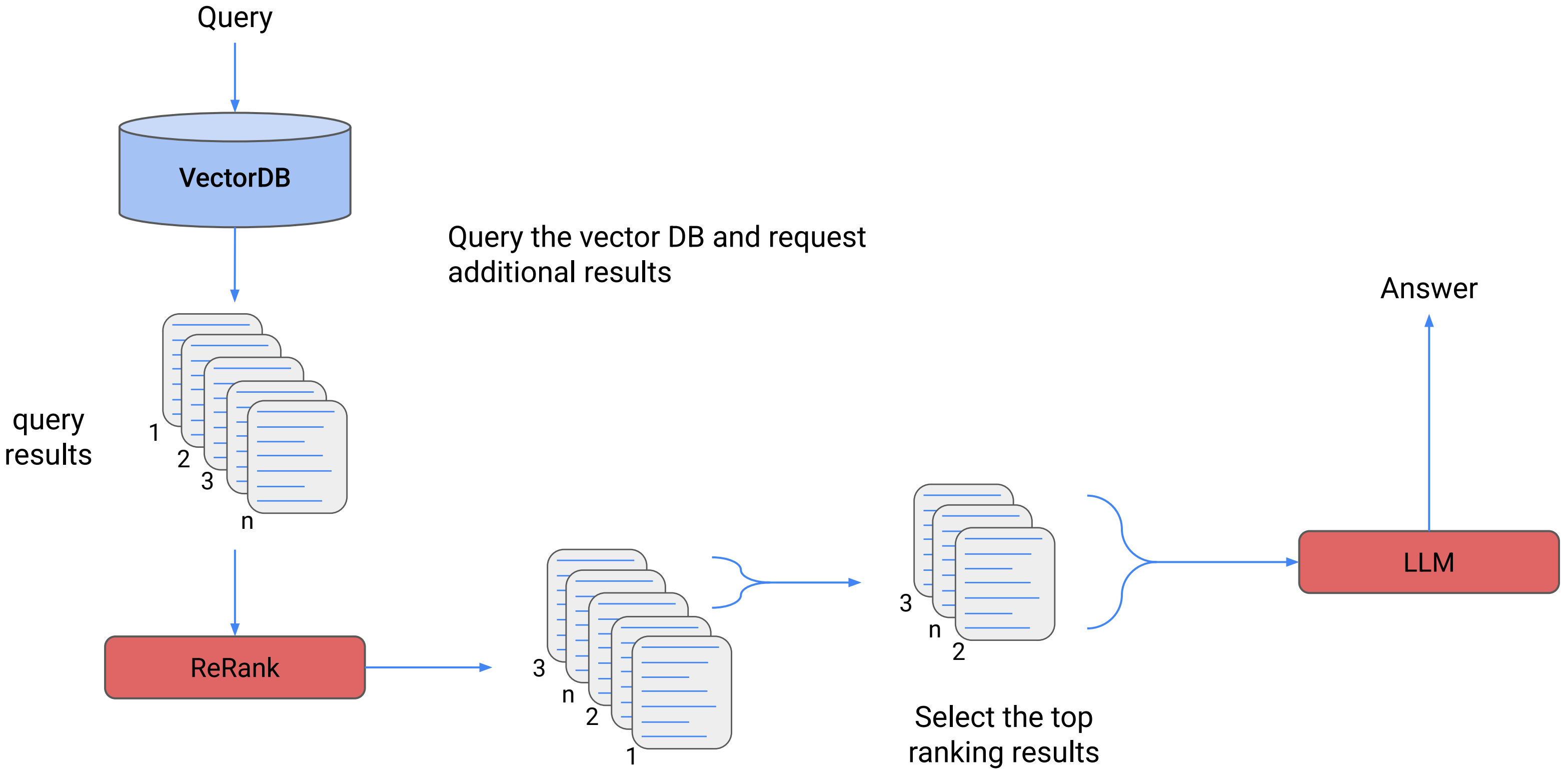


Use an LLM to suggest additional queries

Retrieve results for original and new queries

Send all responses to LLM

ReRanking



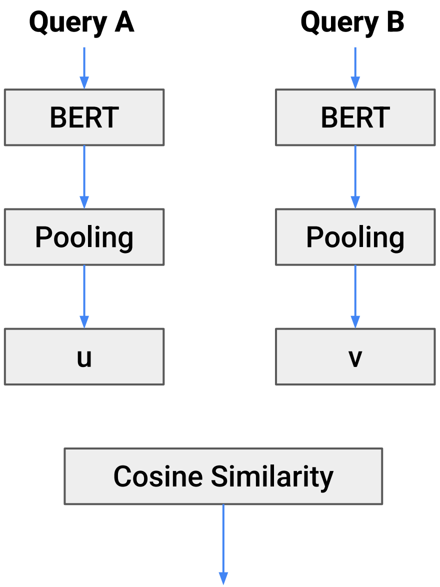
Query the vector DB and request additional results

ReRank output so the most relevant have the highest rank

Select the top ranking results

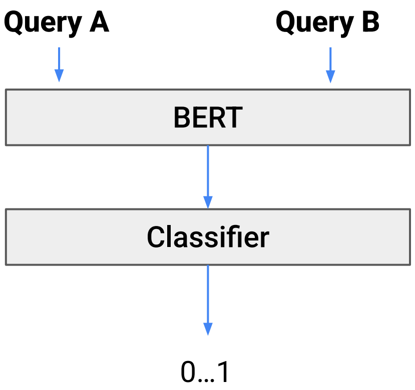
Cross Encoder

Bi-encoder



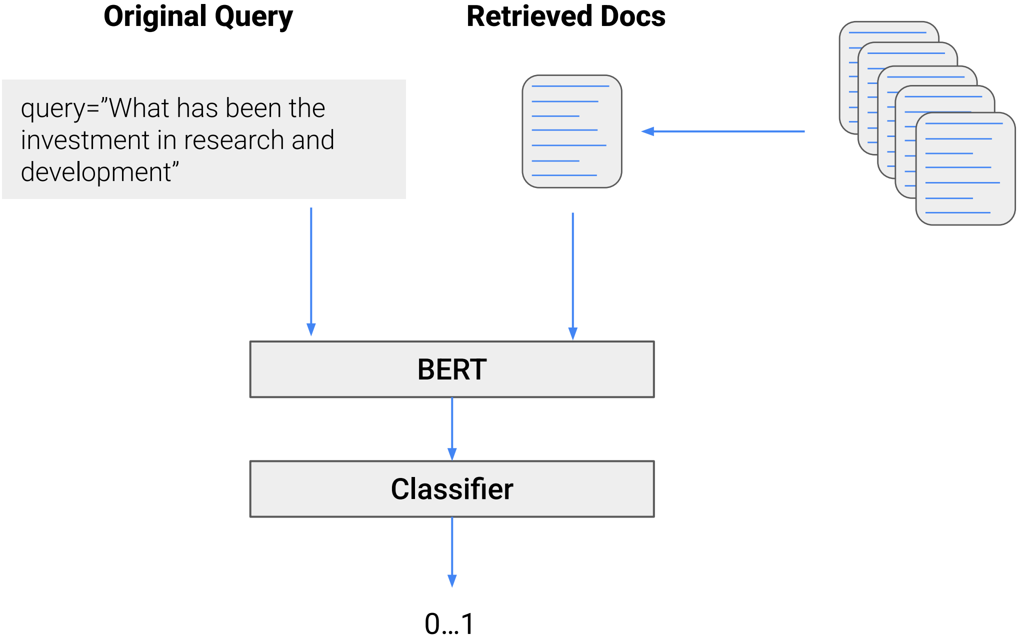
Bi-encoders process two input sequences separately. Each input is fed into its own encoder producing two independent embeddings

Cross-encoder



Cross-encoders process two input sequences together as a single input. This allows the model to directly compare and contrast the inputs, understanding their relationship in a more integrated and nuanced way

Cross Encoder in Re-ranking



Original Query

Retrieved Docs

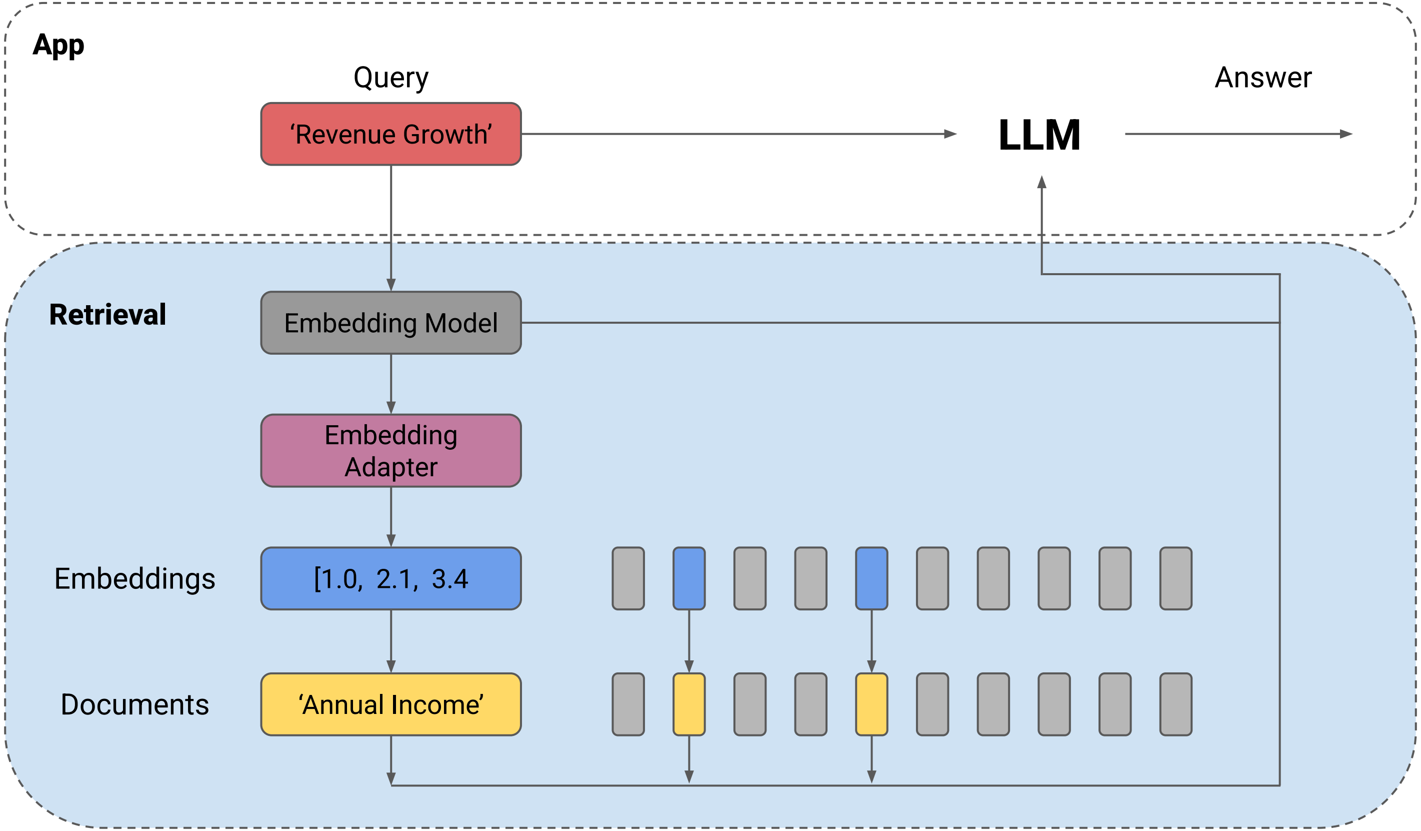
query="What has been the investment in research and development"

BERT

Classifier

0…1

Embedding Adapter



Other Techniques

- Fine-tune the embedding model

- Fine-tine the LLM for retrieval
  - RA-DIT: Retrieval-Augmented Dual Instruction Tuning
  - InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining

- Deep embedding adaptors

- Deep relevance modelling

- Deep chunking

THANK YOU