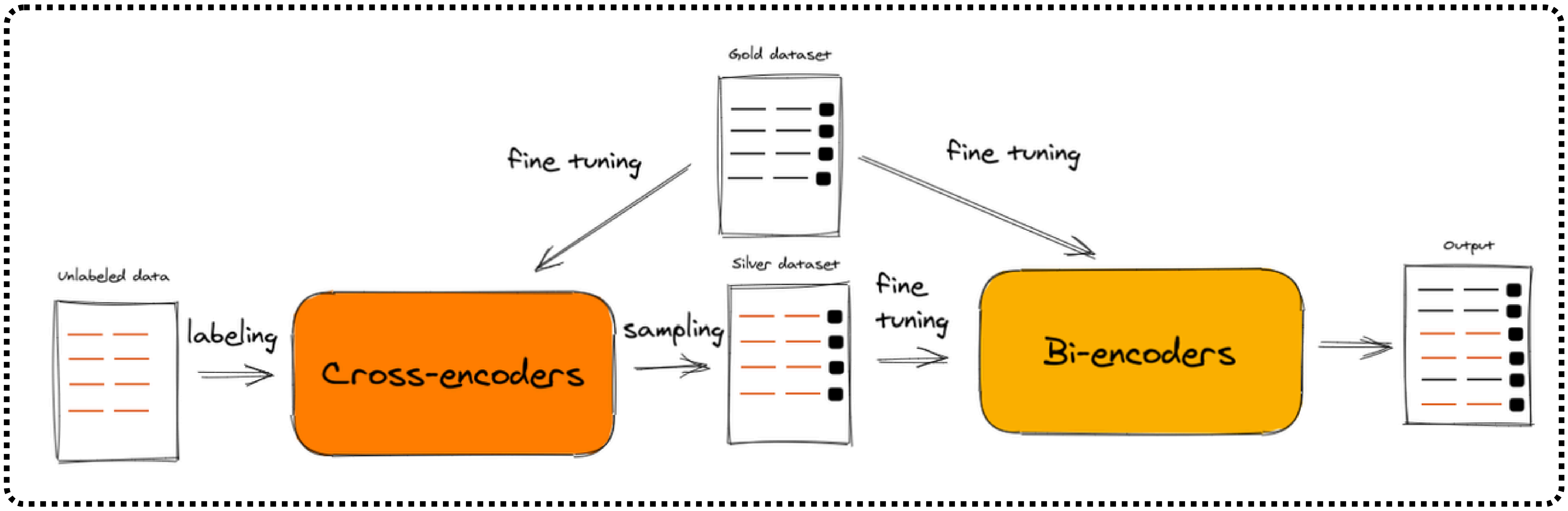
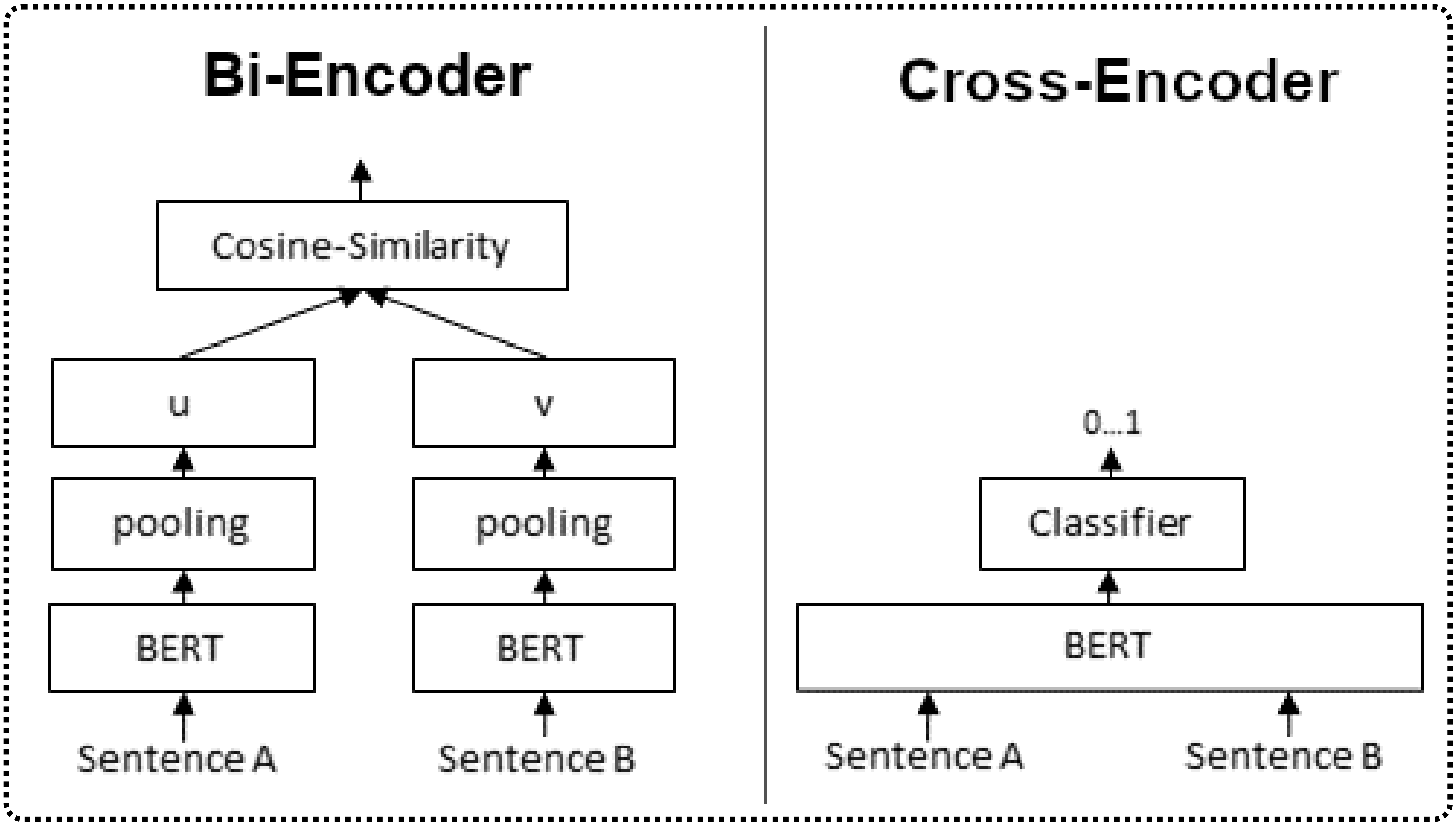
Mastering RAG

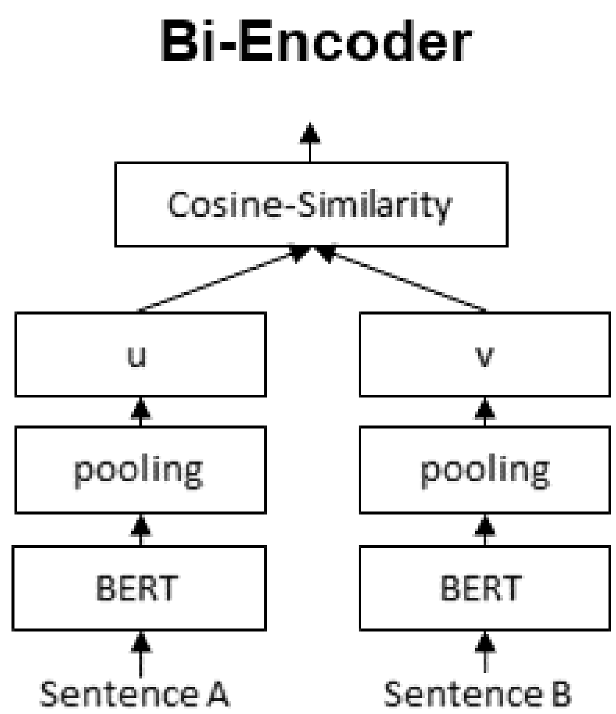
Cross-Encoder vs Bi-Encoder Models in RAG





Bi-Encoder Models

A **bi-encoder** model separately encodes the query and documents using a dual-tower neural architecture, such as a transformer-based encoder (e.g., BERT). The retrieval process follows these steps:



- **Encoding**: The query and each document are independently embedded into a dense vector space.
- **Similarity Computation**: A similarity metric (e.g., cosine similarity, dot product) is used to score document-query pairs.
- **Retrieval**: The top-ranked documents are selected for the generative model.

Advantages

- **Efficiency**: Since embeddings are precomputed for the corpus, retrieval is fast and scalable, especially with Approximate Nearest Neighbors (ANN) techniques like FAISS.
- **Parallelizable**: Queries and documents are encoded independently, making it feasible to process large datasets.
- **Memory Efficient**: Requires storing only the document embeddings, not the entire model.

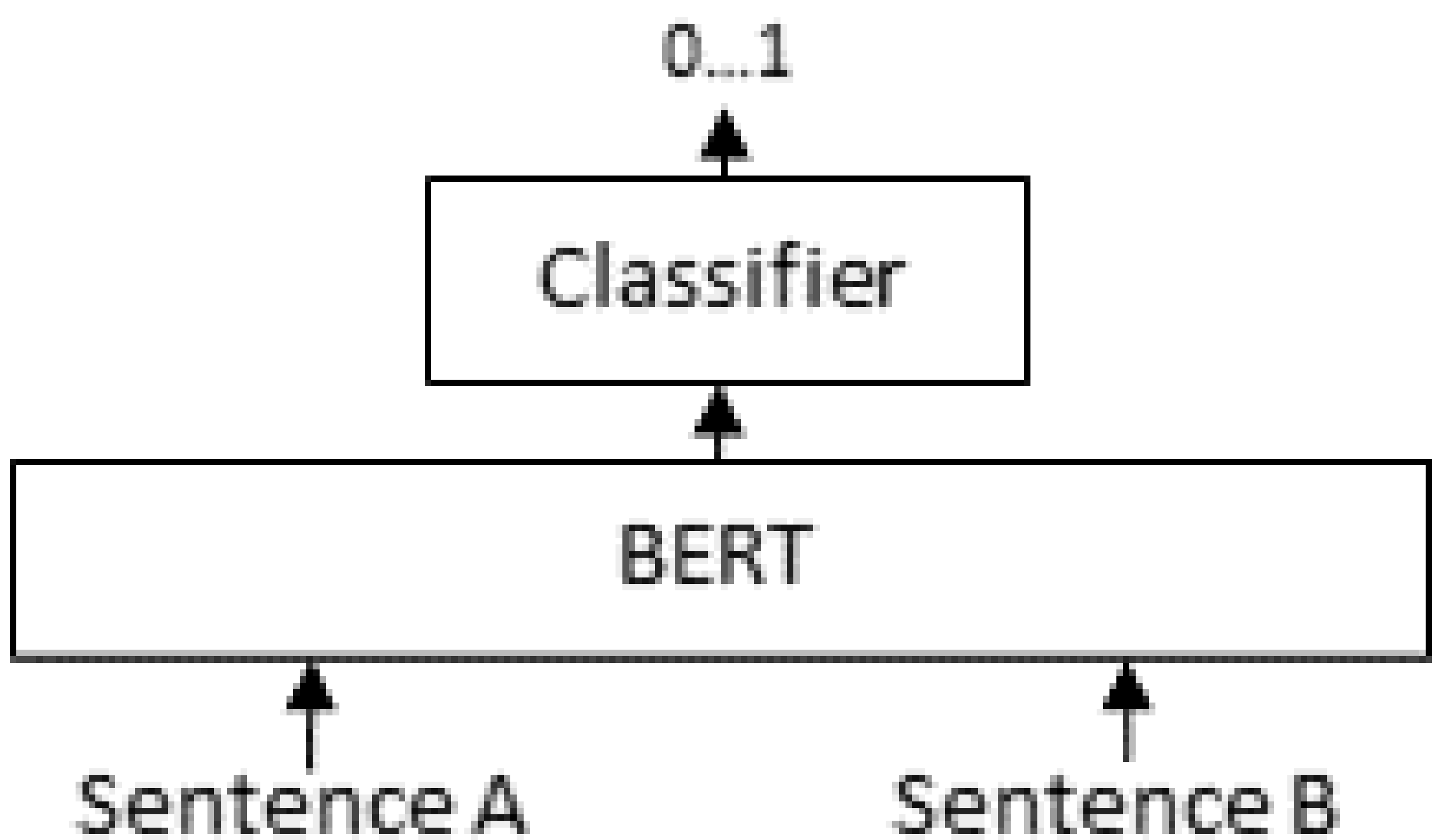
Disadvantages

- **Limited Interaction Modeling**: As the query and document are encoded separately, fine-grained token-level interactions are not captured.
- **Lower Precision**: The retrieval quality may suffer in complex tasks requiring deep semantic understanding.

Cross-Encoder Models

A cross-encoder model processes the query and document together, typically using a transformer-based architecture like BERT. The retrieval process works as follows:

- **Concatenation**: The query and document are combined into a single input sequence.
- **Joint Encoding**: The model jointly processes both inputs, allowing deep attention-based interactions.
- **Scoring**: A classification head or scoring function (e.g., a binary relevance classifier or similarity score) determines the document's relevance.



Advantages

- **Higher Precision:** Since the model processes query-document pairs together, it captures deeper semantic relationships.
- **Better Contextual Understanding**: Token-level attention allows more nuanced ranking decisions

Disadvantages

- **Computational Cost:** Scoring requires running the full model for every query-document pair, making large-scale retrieval infeasible.
- **Not Precomputable:** Unlike bi-encoders, document scores cannot be precomputed, leading to increased latency.

Choosing Between and Cross-Encoder in RAG

Trade-offs in RAG Applications Improved accuracy:

Feature	Bi-Encoder	Cross-Encoder
Speed	High (precomputed embeddings)	Slow (real-time encoding)
Accuracy	Lower (shallow interaction)	Higher (deep interaction)
Scalability	Scalable for large corpora	Limited scalability
Computational Cost	Low (efficient retrieval)	High (real-time processing)