Evaluation of the performance of the retrieval pipeline using metric NDCG@10

**1. Embedding (Small) + Ranking (TinyBERT):**

* **Model Combination**: sentence-transformers/all-MiniLM-L6-v2 + cross-encoder/ms-marco-TinyBERT-L-6
* **Performance**:
  + This combination emphasizes smaller model sizes, resulting in faster retrieval and ranking.
  + **NDCG@10**: Lower than the larger models due to TinyBERT being a smaller, more compressed model. However, it should still provide reasonable performance with efficiency.

**2. Embedding (Large) + Ranking (MiniLM):**

* **Model Combination**: sentence-transformers/all-distilroberta-v1 + cross-encoder/ms-marco-MiniLM-L-12-v2
* **Performance**:
  + Larger embedding model (distilroberta-v1) combined with the more powerful MiniLM-L-12-v2 ranking model.
  + **NDCG@10**: Higher due to the larger capacity of the models, offering a better understanding of semantics and relationships.

**3. Embedding (Small) + Ranking (MiniLM):**

* **Model Combination**: sentence-transformers/all-MiniLM-L6-v2 + cross-encoder/ms-marco-MiniLM-L-12-v2
* **Performance**:
  + A balance between a small embedding model and a more powerful ranking model.
  + **NDCG@10**: Its NDCG@10 score is lies between the previous two. The smaller embedding model limit the accuracy slightly, but the stronger ranking model boost performance during reranking.

**4. Embedding (Large) + Ranking (TinyBERT):**

* **Model Combination**: sentence-transformers/all-distilroberta-v1 + cross-encoder/ms-marco-TinyBERT-L-6
* **Performance**:
  + Large embedding model paired with a smaller ranking model.
  + **NDCG@10**: Its NDCG@10 score is slightly worse than the combination with MiniLM-L-12-v2 but better than the all-small models.

**Overall Comparison:**

* **Best for Accuracy**: **Embedding (Large) + Ranking (MiniLM)** has the highest NDCG@10 score due to the robustness of both models.
* **Best for Efficiency**: **Embedding (Small) + Ranking (TinyBERT)** provides faster performance but with a trade-off in accuracy.
* **Best Balanced**: **Embedding (Small) + Ranking (MiniLM)** offer the best compromise between speed and performance.