

Scratchpad

August 19, 2025

Notes

Section 1

We categorize the models evaluated in this study according to their underlying architecture and training strategies.

Transformer Base Models. As a baseline, we include transformer-based encoder models that have not been further fine-tuned for any specific downstream task. For these models, the final

CLS

token representation is extracted and cosine similarity is used to compute the relevance between the input text and each candidate label. This straightforward approach provides a useful point of reference for subsequent comparisons. The base models considered in this category are the original BERT (*bert-large-uncased* [?]), the increasingly adopted ModernBERT (*ModernBERT-large* [?]), and DeBERTa-v3 (*deberta-v3-large* [?]), a popular and robust modification of BERT that has demonstrated strong performance on a variety of NLP benchmarks.

NLI-based Cross-Encoders. These models are trained on natural language inference (NLI) datasets and perform classification by assessing the degree of entailment between an input text and each candidate label, formulated as a premise–hypothesis pair. *BART-Large-MNLI* [?] is included as the canonical representative, being the first widely used NLI-based cross-encoder for zero-shot classification. We also consider *NLI-RoBERTa-base*, following [?], as well as a set of custom-trained cross-encoders using *BERT*, *DeBERTa-v3*, and *ModernBERT* backbones. Both base and large versions are evaluated to analyze the effect of model scale, and two loss variants are tested to assess the impact of training objectives. Full details of the training procedure are provided in Section ???. In total, 11 NLI-based cross-encoders are benchmarked, covering the most widely used configurations in the literature.

Embedding Models. This category comprises models optimized to produce fixed-size vector representations of text for a range of downstream tasks, including classification. As a canonical embedding model, *all-MiniLM-L6-v2* [?]

is included for its efficiency and strong empirical results, serving as a baseline for this model family. Additionally, we evaluate both base and large variants of BGE, GTE, and E5, all of which use variations of transformer encoders as backbones. To provide contrast, we also include embedding models that leverage large language model architectures, such as Qwen3-Embedding and e5-mistral-7b-instruct; for Qwen3-Embedding, both 0.6B and 8B parameter variants are tested to study the effect of scale. Overall, the embedding model category comprises 11 distinct models.

Rerankers. Reranker models are typically employed in information retrieval, where they re-score candidate documents for relevance to a given query. The *ms-marco-MiniLM-L6-v2* model serves as the reranker counterpart to *all-MiniLM-L6-v2* and is used as the baseline for this group. Similarly, *gte-reranker-modernbert-base* and *bge-reranker-base/large* serve as reranking counterparts to their respective embedding models. We further include *Qwen3-Reranker*, which outputs a relevance score between a document and a query by prompting the model to decide if the document is relevant. The probability assigned to the "yes" token (computed from the model’s vocabulary distribution using a softmax, with all other tokens masked out, except for "yes" and "no") is used as the final relevance score. Both the 0.6B and 8B variants of *Qwen3-Reranker* are evaluated to analyze the impact of model size. In total, 6 reranker models are benchmarked.

Table ?? summarizes the models included in our experiments, listing their architecture, training data, and parameter count. In total, the benchmark covers 29 models.

0.1 Model Types

WE CATEGORIZE THE MODELS EVALUATED IN THIS STUDY ACCORDING TO THEIR UNDERLYING ARCHITECTURE AND TRAINING STRATEGIES.

Transformer Base Models. AS A BASELINE, WE INCLUDE TRANSFORMER-BASED ENCODER MODELS THAT HAVE NOT BEEN FURTHER FINE-TUNED FOR ANY SPECIFIC DOWNSTREAM TASK. FOR THESE MODELS, THE FINAL

CLS

TOKEN REPRESENTATION IS EXTRACTED AND COSINE SIMILARITY IS USED TO COMPUTE THE RELEVANCE BETWEEN THE INPUT TEXT AND EACH CANDIDATE LABEL. THIS STRAIGHTFORWARD APPROACH PROVIDES A USEFUL POINT OF REFERENCE FOR SUBSEQUENT COMPARISONS. THE BASE MODELS CONSIDERED IN THIS CATEGORY ARE THE ORIGINAL BERT (*bert-large-uncased* [?]), THE INCREASINGLY ADOPTED MODERNBERT (*ModernBERT-large* [?]), AND DEBERTA-V3 (*deberta-v3-large* [?]), A POPULAR AND ROBUST MODIFICATION OF BERT THAT HAS DEMONSTRATED STRONG PERFORMANCE ON A VARIETY OF NLP BENCHMARKS.

NLI-based Cross-Encoders. THESE MODELS ARE TRAINED ON NATURAL LANGUAGE INFERENCE (NLI) DATASETS AND PERFORM CLASSIFICATION BY

ASSESSING THE DEGREE OF ENTAILMENT BETWEEN AN INPUT TEXT AND EACH CANDIDATE LABEL, FORMULATED AS A PREMISE–HYPOTHESIS PAIR. *BART-Large-MNLI* [?] IS INCLUDED AS THE CANONICAL REPRESENTATIVE, BEING THE FIRST WIDELY USED NLI-BASED CROSS-ENCODER FOR ZERO-SHOT CLASSIFICATION. WE ALSO CONSIDER *NLI-RoBERTa-base*, FOLLOWING [?], AS WELL AS A SET OF CUSTOM-TRAINED CROSS-ENCODERS USING *BERT*, *DeBERTa-v3*, AND *ModernBERT* BACKBONES. BOTH BASE AND LARGE VERSIONS ARE EVALUATED TO ANALYZE THE EFFECT OF MODEL SCALE, AND TWO LOSS VARIANTS ARE TESTED TO ASSESS THE IMPACT OF TRAINING OBJECTIVES. FULL DETAILS OF THE TRAINING PROCEDURE ARE PROVIDED IN SECTION ?? . IN TOTAL, 11 NLI-BASED CROSS-ENCODERS ARE BENCHMARKED, COVERING THE MOST WIDELY USED CONFIGURATIONS IN THE LITERATURE.

Embedding Models. THIS CATEGORY COMPRISES MODELS OPTIMIZED TO PRODUCE FIXED-SIZE VECTOR REPRESENTATIONS OF TEXT FOR A RANGE OF DOWNSTREAM TASKS, INCLUDING CLASSIFICATION. AS A CANONICAL EMBEDDING MODEL, *all-MiniLM-L6-v2* [?] IS INCLUDED FOR ITS EFFICIENCY AND STRONG EMPIRICAL RESULTS, SERVING AS A BASELINE FOR THIS MODEL FAMILY. ADDITIONALLY, WE EVALUATE BOTH BASE AND LARGE VARIANTS OF BGE, GTE, AND E5, ALL OF WHICH USE VARIATIONS OF TRANSFORMER ENCODERS AS BACKBONES. TO PROVIDE CONTRAST, WE ALSO INCLUDE EMBEDDING MODELS THAT LEVERAGE LARGE LANGUAGE MODEL ARCHITECTURES, SUCH AS QWEN3-EMBEDDING AND E5-MISTRAL-7B-INSTRUCT; FOR QWEN3-EMBEDDING, BOTH 0.6B AND 8B PARAMETER VARIANTS ARE TESTED TO STUDY THE EFFECT OF SCALE. OVERALL, THE EMBEDDING MODEL CATEGORY COMPRISES 11 DISTINCT MODELS.

Rerankers. RERANKER MODELS ARE TYPICALLY EMPLOYED IN INFORMATION RETRIEVAL, WHERE THEY RE-SCORE CANDIDATE DOCUMENTS FOR RELEVANCE TO A GIVEN QUERY. THE *ms-marco-MiniLM-L6-v2* MODEL SERVES AS THE RERANKER COUNTERPART TO *all-MiniLM-L6-v2* AND IS USED AS THE BASELINE FOR THIS GROUP. SIMILARLY, *gte-reranker-modernbert-base* AND *bge-reranker-base/large* SERVE AS RERANKING COUNTERPARTS TO THEIR RESPECTIVE EMBEDDING MODELS. WE FURTHER INCLUDE *Qwen3-Reranker*, WHICH OUTPUTS A RELEVANCE SCORE BETWEEN A DOCUMENT AND A QUERY BY PROMPTING THE MODEL TO DECIDE IF THE DOCUMENT IS RELEVANT. THE PROBABILITY ASSIGNED TO THE "YES" TOKEN (COMPUTED FROM THE MODEL’S VOCABULARY DISTRIBUTION USING A SOFTMAX, WITH ALL OTHER TOKENS MASKED OUT, EXCEPT FOR "YES" AND "NO") IS USED AS THE FINAL RELEVANCE SCORE. BOTH THE 0.6B AND 8B VARIANTS OF *Qwen3-Reranker* ARE EVALUATED TO ANALYZE THE IMPACT OF MODEL SIZE. IN TOTAL, 6 RERANKER MODELS ARE BENCHMARKED.

TABLE ?? SUMMARIZES THE MODELS INCLUDED IN OUR EXPERIMENTS, LISTING THEIR ARCHITECTURE, TRAINING DATA, AND PARAMETER COUNT. IN TOTAL, THE BENCHMARK COVERS 29 MODELS.

Section 2