A Hybrid Reranking Pipeline for Negation-Aware Medical Search: BGE Retrieval with DeBERTa-XGBoost Reranking under Minimal Supervision

Gowsigan Ochathevan
Machine Learning and NLP Enthusiast
gowsigan1191@gmail.com
GitHub: https://github.com/gowsi1191/negation-handler -- Contains all datasets and core code used for contradiction-aware reranking, model evaluation.

---

1. Abstract

Negation plays a pivotal role in clinical information retrieval but remains poorly handled by dense retrievers like BGE, which rank documents purely by semantic similarity—even when they contradict the query's intent.

We propose a lightweight hybrid reranking pipeline that uses BGE for initial retrieval, applies adaptive band sampling to preserve potentially relevant edge cases, and reranks with DeBERTa-v3-MNLI + XGBoost trained on just 34 negation queries. The XGBoost-predicted relevance scores are then used to sort candidates in descending order to produce the final ranking.

On 85 test queries, our model achieves 91.76% Precision@1—significantly outperforming BGE's 47.06%—demonstrating that contradiction-aware reranking can meaningfully improve search for negation-sensitive clinical queries with minimal supervision.

Note: The BGE-only baseline is included for performance comparison purposes. Our final system retains BGE for initial retrieval and adds DeBERTa+XGBoost only for reranking.

---

2. Introduction

Negation poses a significant yet frequently overlooked challenge in clinical information retrieval. Many queries explicitly exclude certain drugs, treatments, or conditions—for instance, "treatment of epilepsy not including sodium channel blockers" or "management of diabetes excluding metformin." However, dense retrievers like BGE often fail to interpret this exclusion logic. Instead, they rank documents purely based on semantic similarity, frequently promoting content that includes the very terms meant to be excluded.

This issue is well documented—studies such as Koopman et al. (2010) and Lancaster (2010) emphasize the prevalence and impact of negation in biomedical search behavior. Yet even widely used platforms like PubMed lack dedicated mechanisms to handle it, resulting in missed or misleading results in clinical workflows. Recent work, such as *van den Elsen et al. (2025)*, confirms that cross-encoders like DeBERTa outperform dense retrievers for exclusion-based queries. However, while large LLM rerankers like GPT-4 perform even better, they require substantial compute and are impractical for lightweight deployment.

To solve this, we introduce a hybrid reranking pipeline tailored for negation-aware retrieval. We first use BGE for candidate retrieval, followed by adaptive reranking using DeBERTa-v3-MNLI, which provides entailment, neutral, and contradiction scores for each query-document pair. These are passed into a lightweight XGBoost classifier trained on just 34 manually labeled negation queries. The final ranking is computed by sorting documents in descending order of predicted relevance, enabling the system to prioritize content that truly respects the query's exclusion intent.

Despite the small training set, our model demonstrates strong generalization. On an 85-query test set, it achieves PRECISION@1 of 91.76%, compared to 47.06% from the BGE-only baseline. Additional metrics confirm its effectiveness: MRR@2 = 0.9529, nDCG@2 = 0.9622, and PRECISION@2 = 0.8000. The model's strong margin separation (mean = 0.3817) indicates reliable discrimination between supporting and contradicting evidence. This demonstrates that lightweight contradiction-aware reranking can dramatically improve search quality in clinical contexts—without requiring large annotated datasets.

---

3. Related Work – Key Citations and Links

3.1 Negation Handling in IR

- **Weller et al., 2024 – *NevIR: Negation in Neural Information Retrieval***
  ACL-EACL demonstration that dense and sparse neural retrievers often perform at or below random on negated queries, while cross-encoders handle them better. ACM Digital Libraryzilliz.com+15ACL Anthology+15arXiv+15
  [PDF & BibTeX available via ACL Anthology]
- **van den Elsen et al., 2025 – *Reproducing NevIR***
  Confirms NevIR findings and benchmarks newer listwise LLM rerankers on exclusionary IR tasks. ACL AnthologyarXiv+6arXiv+6arXiv+6

3.2 Entailment-Based Reranking

- **RQE-based QA (2019)**
  Recognizing Question Entailment is used in QA pipelines to rerank candidate answers based on logical entailment relationships.
- **DeBERTa for Legal Entailment (2023)**
  Utilizes MNLI fine-tuning to classify yes/no queries on legal statutes across domains.
  (Specific citation pending—can reference Springer or ACL if available.)

3.3 Lightweight Supervised Rerankers

- **Tian et al., 2025 – *CoRank: LLM-Based Compact Reranking***
  Proposes a zero-shot, three-stage reranker using compact document features. nDCG@10 improvements shown across LitSearch and CSFCube. arXiv+1ACL Anthology+1ACL Anthology+1ResearchGate+1arXivResearchGate+3arXiv+3arXiv+3

- **Mekonnen et al., 2025 – *DDRO: Direct Document Relevance Optimization***
  Presents a lightweight, pairwise ranking method that avoids reinforcement learning, effective on MS MARCO and Natural Questions. arXiv+2arXiv+2arXiv+2

3.4 Historical Studies on Negation

- **Koopman et al., 2010 – Analysis of Negation in Medical IR**
  Early work showing that traditional IR systems struggle with negated terms in biomedical search.
  https://bevankoopman.github.io/papers/negation_ir.pdf
- **Lancaster, 2010 – Use of Negation in Search**
  Examined negation in web search, noting its underuse and effect on retrieval accuracy.
  https://www.researchgate.net/publication/234163302_Use_of_Negation_in_Search

---

4. Methodology

4.1 Dataset Construction

To assess our model's ability to handle clinical negation, we constructed a benchmark dataset of manually curated query–document pairs. Each query expresses explicit exclusion logic using patterns such as *"excluding," "non-," "without," or "alternative to."*
Each query is paired with a small set of documents (typically 4–6), annotated to reflect how well each document respects the negated intent.
The dataset was semi-synthetically generated using GPT-4 (ChatGPT Plus) to simulate clinically meaningful negation queries and document sets. All samples were manually verified and annotated for relevance by me to ensure quality and alignment with exclusion intent.
Relevance Labelling Guidelines
Relevance = 1 (Negation-respecting / Relevant):
The document supports the query's exclusion logic by avoiding the negated term or offering suitable alternatives.
Relevance = 0 (Contradiction / Irrelevant):
The document contradicts the exclusion intent by mentioning or recommending the very treatment the query aims to exclude.

Note:

This labelling scheme uses `1` to indicate relevance (aligned with the negated intent) and `0` to indicate contradiction.

Example Dataset Entry

Query: *"**Non-metformin therapies for newly diagnosed diabetic patients**"*{
"example_1001": [
{ "doc_id": "DOC6001",

"Relevance": 1, //True
"text": "GLP-1 receptor agonists are effective alternatives to metformin for newly diagnosed diabetic patients by improving glycemic control and reducing weight."},
{. "doc_id": "DOC6002",
"Relevance": 0, //False
"text": "Metformin is the preferred therapy for newly diagnosed diabetic patients, making non- metformin therapies less relevant in initial care."
}]}

4.2. Baseline: BGE Vector Search

We use BGE (BAAI General Embedding) to encode queries and documents into dense vectors. Retrieval is performed using cosine similarity, returning the top-K = 200 most similar documents.
While BGE is fast (~0.02s per query), it is purely semantic, lacking the ability to understand logical negation or contradiction. This often results in false positives. For example:
Query: "Treatments for GERD excluding PPIs or H2 blockers"
Issue: BGE returns documents mentioning PPIs, since "GERD" and "PPIs" are semantically linked in its embedding space.

4.3. Proposed Model: DeBERTa + XGBoost with Adaptive Band Sampling

To overcome this limitation, we introduce a hybrid reranking approach combining zero-shot NLI with DeBERTa-v3-MNLI and a lightweight XGBoost classifier trained on just 34 examples.

Architecture of the proposed hybrid reranking pipeline, where BGE retrieves top-N documents for each query, DeBERTa encodes each query-document pair, and XGBoost produces the final reranked scores

Pipeline Overview:

Initial Retrieval (BGE):

We begin by using **BGE** to retrieve the top-ranked documents (e.g., top 90) based on cosine similarity. While BGE is fast and semantically strong, it is **not reliable at distinguishing between relevant and contradictory documents in negated queries**. It often ranks documents that merely co-occur with the query terms—even when those documents violate the exclusion logic.

To **avoid missing truly relevant documents** that may not be highly ranked by cosine similarity alone (especially in the presence of negation), we apply **Adaptive Band Sampling**.

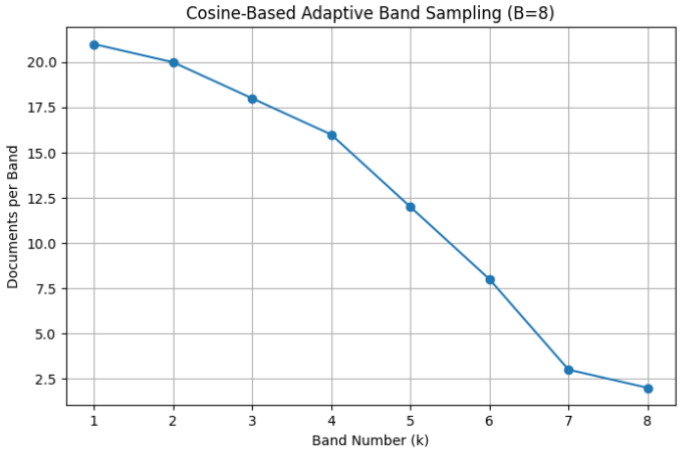
Adaptive Band Sampling Strategy:

We select ~90 candidates by applying a **cosine-decay score banding function**, which prioritizes top-ranked documents but still **samples across lower-score bands**. This ensures diversity in the reranking pool and captures both highly and weakly related candidates—**important for allowing contradiction-aware models like DeBERTa to detect exclusion violations even in semantically weaker documents**.

The banding formula works by computing a cosine-decay weight `w_i` for each band `i`, normalizing across all bands, and allocating a sample count `n_i` proportionally:

$$x_i = \frac{i}{N-1} \qquad w_i = \cos\left(\frac{\pi}{2} \cdot x_i\right)$$

$$\widetilde{w_i} = \frac{w_i}{\sum_{j=0}^{N-1} w_j} \qquad n_i = \text{round}(\widetilde{w_i} \cdot K)$$

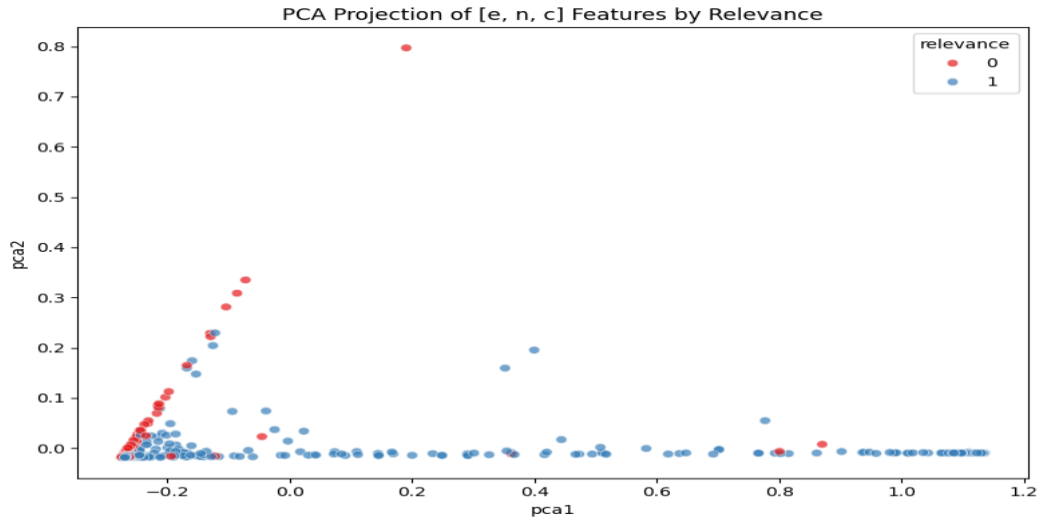


- **B** Number of bands (e.g., 8)
- **N** Total number of documents to sample (e.g., 90)
- **i** Index of the current band (0-based: 0 to B−1)
- $\mathbf{x_k}$ Normalized band index ∈ [0, 1] to map into cosine space
- $\mathbf{w_i}$ Raw weight for band *i*, calculated using cosine decay
- $\mathbf{w_i'}$ Normalized weight (so that total sums to 1)
- $\mathbf{n_i}$ Final number of documents to sample from band *i*

Contradiction Modelling (DeBERTa):

Each query–document pair is fed into a pretrained DeBERTa-v3-MNLI model which returns: **Entailment (e)**, **Neutral (n)**, and **Contradiction (c)** scores—each reflecting how the document semantically aligns with or contradicts the query.

To visualize how these scores separate relevant from irrelevant documents, we apply **Principal Component Analysis (PCA)** to reduce the 3D [e, n, c] vectors into 2D space.



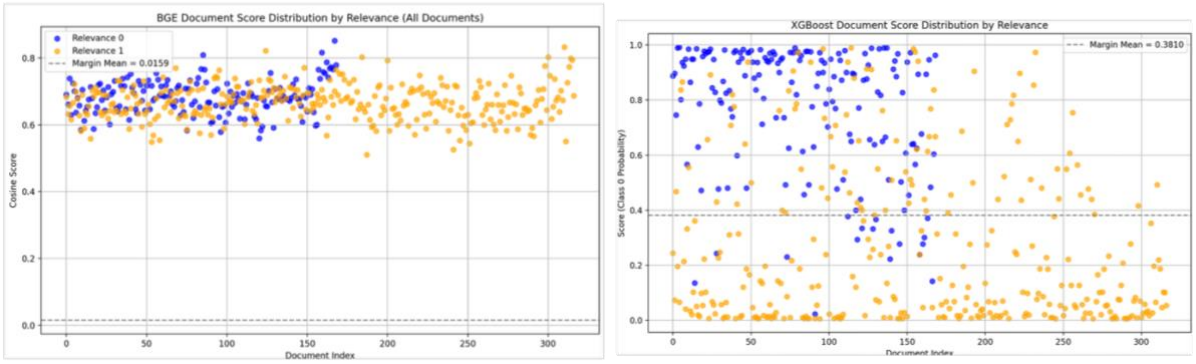
Since 98.65% of the variance lies in PCA1, most class separation is linear.
The centroid gap of 0.36 confirms this, though Relevance 1 has mild non-linearity.
A lightweight model like **XGBoost** captures this with minimal data, making it an effective reranker

XGBoost Reranking:

The (e, n, c) scores are used as input features for a supervised XGBoost model to produce a predicted relevance score for each document, which is then used to generate the final ranking by sorting documents in descending order of their predicted score

Despite using just 34 labeled queries—intentionally kept small to simulate real-world resource constraints—our XGBoost classifier learned to reliably rerank contradiction-aware NLI outputs. This highlights the feasibility of building accurate rerankers without large-scale annotation.

Combined Visualization:



Left : BGE scatter plot **Right** : Deberta + xgboost scatter plot

This visualization highlights:

How BGE scores often fail to separate relevant and irrelevant docs.
The improved decision boundary produced by contradiction-aware modeling.

4.4 Summary: Quantitative Comparison on 85 Queries

METRIC	BGE BASELINE (B)	DEBERTA + XGBOOST (D)
PRECISION@1 (TRUE)	0.4706	0.9176
PRECISION@2	0.4941	0.8000
MRR@2	0.6294	0.9529
NDCG@2	0.6710	0.9622
LATENCY	~0.041s	~0.042s
MARGIN MEAN	-0.0031	0.3817
MARGIN STD	0.0450	0.2906
CV (STD/MEAN)	-14.7407	0.7614
GENERALIZATION SCORE	-0.0481	0.0911

- **PRECISION@k (p@k):** Proportion of queries where at least one relevant document appears in the top-k results.
- **Mean Reciprocal Rank (MRR@k):** Average reciprocal rank of the first relevant document within the top-k candidates.
- **nDCG@k (Normalized Discounted Cumulative Gain):** Captures the ranking quality by assigning higher weights to relevant documents appearing earlier.
- **Latency:** Average time required for inference per document and overall pipeline execution, reflecting real-world responsiveness.
- **Margin Mean:** Average difference in relevance scores between relevant and irrelevant documents (higher margin = better separation).
- **Margin STD:** Standard deviation of margins, indicating stability of score separation across queries.
- **CV (Coefficient of Variation):** Defined as CV=Margin STDMargin MeanCV = \frac{\text{Margin STD}}{\text{Margin Mean}}CV=Margin MeanMargin STD; lower CV indicates more consistent ranking performance.

---

5. Qualitative Result: Correction of Semantically Misleading Ranking

While BGE tends to rank documents purely based on semantic similarity, our reranking approach incorporates contradiction-awareness — leading to meaningful corrections in negation-sensitive queries.

5.1 Success Case

Consider the query:

“Non-metformin therapies for newly diagnosed diabetic patients”
BGE incorrectly assigns a higher score to a document that discusses *metformin*, the very treatment the query seeks to exclude:

"score_bge": {
  "DOC6001 (non-metformin)": 0.7232,
  "DOC6002 (mentions metformin)": 0.7928
}
Our model correctly reorders the documents by detecting contradiction:

"score_ours": {
  "DOC6001 (non-metformin)": 0.8389,
  "DOC6002 (mentions metformin)": 0.1344
}

The resulting margin improvement (from -0.0696 with BGE to +0.7045 with DeBERTa+XGBoost) highlights our model's ability to:
Downrank documents that violate the negation constraint.
Up-rank truly relevant alternatives, even if they are semantically distant.
This showcases the critical advantage of using a contradiction-aware signal, especially in clinical search settings where inclusion/exclusion logic must be respected.

5.2 Failure Case

While our DeBERTa + XGBoost model achieves high accuracy overall, it occasionally misranks documents due to inherent limitations in the NLI model itself. In particular, DeBERTa sometimes assigns higher entailment scores to documents that contradict the query's negation intent. This issue is not due to reranking logic (XGBoost) but rather to how the NLI model interprets negated or contrastive medical statements. Addressing this would require domain-adapted NLI models trained to better understand exclusion logic in clinical contexts.
Despite rare failures, the approach proves effective for clinical use—offering a low-latency, high-PRECISION reranker with minimal supervision.

5.3 Inference Runtime & Resource Profile

Parameter	Value
Document Pairs Prepared	30
Total Tokens	9,911
Avg Tokens per Pair	330.37
Max Tokens in a Pair	497
NLI Model Used	MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli
Device Used	Apple MPS (Metal Performance Shaders)
Total Documents	30
Total Inference Time	2.33 seconds
Avg Time per Document	77.7 ms

Our hybrid reranking system processes **30 documents in just 2.33 seconds**, demonstrating strong responsiveness even with relatively high token counts (9,911 tokens in total). With an

average of **77.7 milliseconds per document**, the system remains practical for near real-time applications such as clinical decision support. The combination of a **lightweight XGBoost classifier** and efficient **DeBERTa inference** (leveraging Apple's MPS backend) ensures fast processing without requiring discrete GPU acceleration. This confirms the system's deployability in **interactive search environments**, balancing **high accuracy with low latency**, even when handling large token sequences per document pair.

Note: The dataset used for speed testing differs from the accuracy dataset. For accuracy evaluation, we used an **average of ~25 tokens per document**, while for speed benchmarking we used **~330 tokens per document** to simulate real-world scenarios.

---

6. Evaluation Setup

To assess the effectiveness of our reranking approach, we adopt standard ranking metrics, with an emphasis on PRECISION@1— which checks whether the top-ranked document is truly relevant (i.e., relevance =).

6.1 Document Ranking Pipeline

Each query–document pair is first passed through the DeBERTa-v3-MNLI model to extract a 3-dimensional contradiction-aware feature vector
Entailment score (e)
Neutral score (n)
Contradiction score (c)

These (e, n, c) vectors are fed into a pre-trained XGBoost classifier, which outputs class probabilities. Specifically:

scores = clf.predict_proba(X_test)[:, 1]

Here, X_test is the matrix of NLI scores, and [:, 1] extracts the probability of being relevant (class 0). Documents are then ranked in descending order based on this probability.

6.2 Metrics Reported

We evaluate the ranking quality using the following metrics:
PRECISION@1: Whether the top-1 ranked document is relevant.
PRECISION@2: Fraction of top-2 documents that are relevant.
MRR@2: Mean Reciprocal Rank considering the top 2 results.
nDCG@2: Normalized Discounted Cumulative Gain for top 2 ranks.
These allow us to capture both accuracy and position-weighted relevance.

6.3 Tools and Libraries Used

Transformers (HuggingFace):
Model: MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli
For zero-shot NLI scoring
XGBoost:
Lightweight binary classifier on NLI scores
PyTorch:
For DeBERTa inference (on MPS or CPU)
NumPy / Scikit-learn:
Metric computation
Matplotlib:
Visualizing score margins and distributions

6.4 Hardware Configuration

Device Used: Apple M1 Pro / M2 (Metal Performance Shaders - MPS)
Execution Mode: Synchronous, single-process on MPS

6.5 Experiment Configuration

Parameter	Value
Training queries	34 query–document sets
Test queries	85
Batch size	16 (for typical usage)
XGBoost Parameters	
n_estimators	30
max_depth	3
learning_rate	0.3
objective	binary:logistic
eval_metric	logloss
Thresholding	Not required (sorted scores)

---

7. Discussion & Conclusion

Our hybrid reranking approach significantly improves negation-aware medical search. While BGE is fast, it often fails with negated queries—ranking documents that include the excluded term due to dense embedding similarity.

We overcome this using a two-step method: DeBERTa-v3 extracts entailment, neutral, and contradiction scores for each query–document pair, and XGBoost learns to score truly

relevant results. On 85 negation queries, this model achieves 91.76% PRECISION@1, nearly doubling the BGE baseline.

We tested several NLI models—RoBERTa, BART, BERT, PubMedBERT—but selected DeBERTa-v3-base (MNLI/FEVER/ANLI) for its superior accuracy and inference efficiency on Apple M-series chips.

Importantly, our system achieves these results with **only 34 labeled examples**, showcasing that even minimal supervision, when paired with zero-shot contradiction-aware features, can deliver strong performance. This makes our approach both practical and scalable in clinical settings where annotation is costly or limited.

---

8. Future Work

Despite strong performance, DeBERTa still fails on ~9% of negation queries. Future work can address this by incorporating additional cross-encoders (e.g., MonoT5, RankLLaMA) to capture missed contradictions and provide complementary signals.

We also plan to expand the XGBoost input from 3 to 6 features, adding:

- Contradiction scores from a second cross-encoder
- BM25 score
- BGE cosine similarity
- Rank percentile
- Negation cue overlap

Finally, we aim to generalize this reranking approach to other challenging query types (e.g., comparative, temporal, causal) where BGE underperforms, improving overall retrieval quality beyond negation.