

SIREN: Improving Systematic Review Screening with Synthetic Training Data for Neural Retrievers

Praise Oketola

praise.o.dev@gmail.com

Abstract

Systematic reviews require screening thousands of documents against detailed eligibility criteria, a process consuming 40–80 hours of expert time per review. While biomedical dense retrievers like MedCPT have advanced PubMed search, we find they *underperform* recent general-domain models on criteria-based screening, exhibiting larger performance gaps on partial-relevance queries. We present SIREN, a retrieve-then-rerank system trained on synthetic data with explicit graded relevance. Our pipeline generates $\sim 60,000$ query-document pairs using a large language model (LLM) teacher with seven diversity dimensions, enabling training without manual annotation. On external benchmarks, SIREN matches WSS@95 on CLEF-TAR 2019 while improving ranking quality by +12.0pp NDCG@10, and outperforms baselines on 5 of 7 comparisons across CLEF-TAR, CSMed, and BEIR. On our released screening benchmark, SIREN achieves 0.94 MRR@10, a +24 percentage point (pp) improvement over MedCPT. We release SIREN-60k with canonical splits to support reproducible research in systematic review automation.

1 Introduction

Systematic reviews synthesize evidence from primary studies to answer specific research questions, forming the backbone of evidence-based medicine (Higgins et al., 2019). The screening phase requires determining which studies meet predefined inclusion criteria, a task that dominates the review timeline. A typical review screens 5,000–10,000 abstracts, with some exceeding 50,000 (Borah et al., 2017). At an estimated 30 seconds per abstract, screening alone can consume 40–80 hours of expert time.

Current automation tools like ASReview (van de Schoot et al., 2021) use active learning with classical features (TF-IDF, document embeddings) to

prioritize likely-relevant documents. While effective at reducing screening burden, these approaches have limitations: (1) they require initial seed documents, (2) they optimize for binary relevance despite evidence that relevance is graded (Sormunen, 2002), and (3) they do not leverage recent advances in dense retrieval.

One might expect biomedical-specific dense retrievers to excel at this task. MedCPT (Jin et al., 2023), trained on 255 million PubMed query-article pairs, represents the state of the art in biomedical retrieval. Yet our analysis reveals a surprising pattern: biomedical models exhibit *larger* performance gaps between fully-relevant and partially-relevant queries (24–28pp) than general-domain dense retrievers (18–24pp). This suggests the challenge is not domain vocabulary but rather the *training signal*: models trained on keyword-style search logs lack the graded relevance signal needed to distinguish documents that match all inclusion criteria from those that match only some.

We present SIREN (Systematic Information Retrieval and Evidence Navigation), a neural screening system that addresses this gap through synthetic data generation with explicit graded relevance. Our contributions are:

1. A **retrieve-then-rerank pipeline** that improves NDCG@10 by +12.0pp over MedCPT on CLEF-TAR 2019 and leads on 5 of 7 external comparisons. On our released benchmark, the pipeline achieves +24pp MRR@10 over MedCPT (0.94 vs. 0.70).
2. A **synthetic data generation framework** with seven diversity dimensions that produces 60,000 training examples without manual annotation. We show that graded relevance training reduces the positive-partial performance gap from 18pp (best general-domain baseline) to 10pp (45% reduction).

3. **SIREN-60k**, a benchmark for criteria-based screening with canonical train/val/test splits and bootstrap confidence intervals, enabling reproducible evaluation of future work.
4. **Comprehensive analysis** revealing that biomedical-specific models underperform general-domain models on this task, and that explicit 3-class relevance training (Relevant/Partial/Irrelevant) enables practical triage workflows.

2 Related Work

2.1 Systematic Review Automation

Technology-assisted review (TAR) emerged from legal e-discovery and was adapted for systematic reviews through the CLEF-TAR shared tasks (Kanoulas et al., 2017, 2019). The standard metric, Work Saved over Sampling at 95% recall (WSS@95), measures the fraction of documents a reviewer can skip while maintaining near-complete recall (Cohen et al., 2006). Continuous Active Learning (CAL), which trains classifiers iteratively as reviewers label documents, has become the dominant paradigm for high-recall retrieval (Cormack and Grossman, 2014, 2016). Modern tools include ASReview (van de Schoot et al., 2021), Rayyan (Ouzzani et al., 2016), and Abstrackr (Wallace et al., 2012). These systems primarily use active learning with sparse features (TF-IDF, word embeddings). Recent work explores dense retrieval for initial ranking (Mao et al., 2024), but focuses on ranking rather than the retriever training data itself.

2.2 Biomedical Neural Retrieval

Domain-specific pretrained models like BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2022), and SciBERT (Beltagy et al., 2019) provide biomedical language understanding but are not trained for retrieval. MedCPT (Jin et al., 2023) addresses this gap by training contrastive encoders on 255 million query-article pairs from PubMed search logs. BioLORD (Remy et al., 2024) uses ontology-guided pretraining for biomedical sentence similarity. However, these models are trained on keyword-style search queries (e.g., “diabetes treatment”), which differ substantially from structured eligibility criteria (e.g., “RCTs of metformin in type 2 diabetics over 65 with HbA1c > 7%”). We build on GTE (Li et al., 2023), a

general-purpose encoder with strong zero-shot performance, and demonstrate that task-specific fine-tuning with appropriate training data outperforms biomedical-specific pretraining.

2.3 Synthetic Data for Information Retrieval

Generating synthetic queries for retrieval training has proven effective when labeled data is scarce. InPars (Bonifacio et al., 2022) uses few-shot prompting to generate queries, applying aggressive filtering to remove low-quality examples. GPL (Wang et al., 2022) combines query generation with cross-encoder pseudo-labeling for domain adaptation. Promptagator (Dai et al., 2023) achieves strong results with only 8 task-specific examples by leveraging large language models. These approaches use *binary* relevance (relevant/irrelevant); we extend this paradigm with explicit *graded* relevance labels, enabling models to distinguish fully relevant documents from partial matches, a critical capability for systematic review triage. We also introduce diversity controls specific to eligibility criteria, including PICO (Population, Intervention, Comparison, Outcome) structure and exclusion criteria with negation.

3 Method

3.1 Problem Formulation

Given a query q describing eligibility criteria for a systematic review (e.g., “RCTs of aspirin for cardiovascular prevention in diabetic patients”) and a corpus $\mathcal{D}$ of biomedical abstracts, the task is to rank documents by relevance. Following work on graded relevance (Sormunen, 2002; Kekäläinen, 2005), we adopt three-class relevance:

- **Relevant:** Matches ALL core criteria
- **Partial:** Matches ≥ 1 but not all core criteria
- **Irrelevant:** Matches zero core criteria

This enables triage workflows where reviewers prioritize fully relevant documents. Table 1 illustrates these distinctions with a concrete example.

3.2 Architecture

SIREN uses a two-stage retrieve-then-rerank architecture:

Query	Randomized controlled trials evaluating probiotics in adult patients with irritable bowel syndrome, measuring symptom severity with validated scales
Relevant	A double-blind RCT of <i>Lactobacillus rhamnosus</i> in 120 adult IBS patients showed significant reduction in IBS-SSS scores compared to placebo ($p < 0.001$). (All criteria met: RCT design, probiotics, adult population, IBS condition, validated outcome scale)
Partial	An observational cohort study of probiotic supplementation in 200 adult IBS patients showed improvement on IBS-SSS scores over 12 weeks. (Most criteria met: probiotics, adult population, IBS, validated scale; Missing: RCT design)
Irrelevant	This narrative review discusses the role of gut microbiome composition in various gastrointestinal disorders including inflammatory bowel disease. (No criteria met: review not RCT, no probiotics intervention, no IBS, no validated outcomes)

Table 1: Example of 3-class relevance classification. A document is **Relevant** if it matches all eligibility criteria, **Partial** if it matches some but not all criteria, and **Irrelevant** if it matches none.

Stage 1: Bi-encoder Retrieval. Following the dual-encoder paradigm (Karpukhin et al., 2020; Reimers and Gurevych, 2019), we encode queries and documents independently using a shared encoder E :

$$\text{score}(q, d) = E(q)^\top E(d) \quad (1)$$

This enables efficient nearest-neighbor search over large corpora.

Stage 2: Cross-encoder (CE) Reranking. Top- k candidates are reranked using a cross-encoder (Nogueira and Cho, 2019) that jointly encodes query-document pairs:

$$\text{score}(q, d) = \text{MLP}(\text{CLS}([q; d])) \quad (2)$$

The cross-encoder outputs three-class probabilities. For ranking, we compute a weighted score: $P(\text{Partial}) + 2 \cdot P(\text{Relevant})$, which ensures fully relevant documents rank above partial matches.

3.3 Synthetic Data Generation

We generate training data using a teacher LLM (DeepSeek-v3.2)¹ prompted with diverse biomedical articles from PubMed. For each article, the teacher generates: (1) positive queries where the article fully matches all criteria, (2) partial queries where the article matches some but not all criteria, (3) synthetic hard negative articles (500–600 words) that share vocabulary and topic but explicitly violate key criteria, and (4) similar titles for title-to-title training pairs.

To ensure diversity in generated queries, we control seven dimensions via configurable “knobs”:

1. **Length:** Short (1–2 sentences), medium (3–5 sentences), long (5+ sentences)
2. **Format:** Prose, bulleted list, PICO-structured, Boolean search syntax
3. **Specificity:** Broad (general topic), moderate, narrow (highly specific)
4. **Criteria count:** Single criterion, few (2–3), many (4+)
5. **Terminology:** Formal medical terms, lay language, mixed
6. **Negation:** None, simple exclusions, complex exclusions
7. **Ambiguity:** Clear criteria, moderate ambiguity, underspecified

Each dimension has weighted sampling probabilities calibrated to reflect real-world query distributions (e.g., 80% short queries, 88% prose format). Incompatible combinations (e.g., short length with many criteria) are automatically resampled. Figure 1 illustrates our data generation pipeline.

Negative Sampling. Each training example pairs a query with its source article (positive) and hard negatives. We use a 70/20/10 sampling strategy: 70% teacher-generated negatives (articles that explicitly violate specific criteria), 20% BM25-retrieved negatives (lexically similar but semantically irrelevant), and 10% random negatives (for distribution diversity). The teacher-generated negatives are particularly valuable as they capture subtle criteria mismatches that lexical methods miss.

¹<https://github.com/deepseek-ai/DeepSeek-V3>

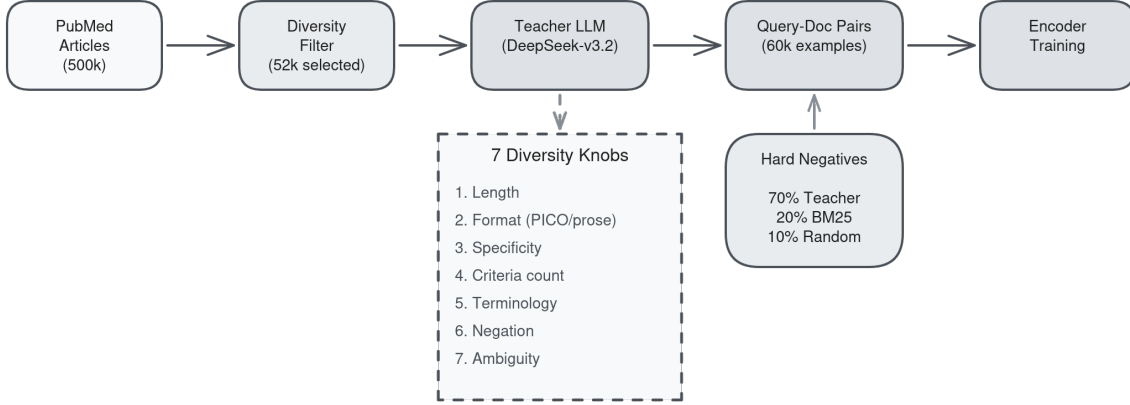


Figure 1: SIREN data generation pipeline. PubMed articles are filtered for diversity, then processed by a teacher LLM with configurable diversity knobs to generate varied query-document pairs. Hard negatives are sampled from teacher-generated (70%), BM25-mined (20%), and random (10%) sources.

3.4 Training

Bi-encoder. We fine-tune GTE-ModernBERT-base (Li et al., 2023; Warner et al., 2024) using Multiple Negatives Ranking Loss (MNRL) with gradient caching (Gao and Callan, 2021), which enables large effective batch sizes without proportional memory increase. With batch size 512, each query sees 511 in-batch negatives plus 7 explicitly sampled hard negatives. We train for 1 epoch with learning rate 2×10^{-5} and Matryoshka Representation Learning (Kusupati et al., 2022) at dimensions (768, 512, 256, 128, 64) for flexible deployment. We apply data augmentation during training: 15% title dropout (randomly removing the article title) and 10% title swap (replacing the title with a different example’s title), improving robustness to title quality variation.

Cross-encoder. We fine-tune GTE-reranker-ModernBERT-base for three-class classification (Relevant=2, Partial=1, Irrelevant=0) using cross-entropy loss. The final relevance score is computed as $P(\text{Partial}) + 2 \cdot P(\text{Relevant})$, which ranks fully relevant documents above partial matches.

Model Merging. A common failure mode of fine-tuning is catastrophic forgetting: task-specific gains come at the cost of out-of-domain (OOD) performance. We observe this empirically: the fine-tuned bi-encoder (before merging) gains +5pp on our internal benchmark but degrades performance on external benchmarks compared to the base model.

Following Wortsman et al. (2022b,a), we address this via SLERP (Spherical Linear Interpolation) merging, which interpolates between fine-

tuned and base model weights. For the bi-encoder, $t = 0.4$ (40% fine-tuned, 60% base) recovers OOD performance while preserving task gains. For the cross-encoder, we use $t = 0.2$ due to its higher sensitivity to weight perturbation. The merged bi-encoder achieves +24pp over MedCPT on SIREN (+7.6pp over its own base model) while maintaining or exceeding GTE-ModernBERT-base on external benchmarks.

4 Experiments

4.1 Datasets

SIREN-60k (Internal). ~60,000 synthetic query-document pairs split into 56,950 train / 1,500 val / 1,500 test documents. Each document yields multiple queries (positive and partial), resulting in approximately 9,000 test queries, balanced between positive and partial types. As the test set is drawn from the same generation pipeline, we treat it as an intrinsic evaluation of model fit and rely on CLEF-TAR, CSMed, and BEIR for generalization assessment.

CLEF-TAR 2019. Systematic review screening benchmark using the Intervention subtask (20 topics), the standard evaluation subset for treatment effectiveness reviews (Kanoulas et al., 2019).

BEIR Biomedical. We evaluate on SciFact, TRECCOVID, and NFCorpus from the BEIR benchmark (Thakur et al., 2021).

4.2 Baselines

We compare against:

- **BM25:** Sparse retrieval baseline

- **MedCPT** (Jin et al., 2023): Biomedical-specific dense retriever
- **GTE-ModernBERT-base** (Li et al., 2023; Warner et al., 2024): General-purpose dense retriever (SIREN’s base model)
- **BGE-base-en-v1.5** (Xiao et al., 2023): General-purpose dense retriever
- **EmbeddingGemma-290m** (Vera et al., 2025): General-purpose embeddings (290M params)
- **PubMedBERT-base-embeddings²** (Gu et al., 2022): Biomedical encoder (mean pooling)
- **BioLORD-2023** (Remy et al., 2024): Ontology-guided biomedical embeddings

4.3 Metrics

For retrieval, we report MRR@10, NDCG@10 (Järvelin and Kekäläinen, 2002), and Recall@k. For systematic review screening, we report WSS@95 (Work Saved over Sampling at 95% recall), the standard metric for TAR evaluation (Cohen et al., 2006). All internal benchmark results include bootstrap 95% confidence intervals from 1,000 resamples.

5 Results

5.1 Internal Benchmark

Table 2 presents results on the SIREN test set, comparing our models against biomedical-specific retrievers (MedCPT, BioLORD, PubMedBERT), general-purpose dense retrievers (GTE, BGE, EmbeddingGemma), and sparse retrieval (BM25). All neural models use base-sized architectures (~ 110 M parameters) for fair comparison, except EmbeddingGemma (290M).

SIREN achieves 0.937 MRR@10, outperforming MedCPT (0.697) by 24.0pp with non-overlapping 95% confidence intervals ([0.93–0.94] vs. [0.69–0.71]). Adding the cross-encoder yields a further 1.5pp improvement to 0.952 MRR@10. Figure 2 visualizes these results. External benchmarks (Section 5.2) provide generalization evidence, with SIREN leading in 5 of 7 comparisons across CLEF-TAR, CSMeD, and BEIR, and tying a sixth.

Biomedical-specific models (MedCPT, BioLORD, PubMedBERT) *underperform* general-domain models (GTE, BGE) on this task. MedCPT ranks lowest among neural models despite

Model	MRR@10	R@10	NDCG@10
<i>Biomedical Models</i>			
MedCPT-base	0.697 (.69–.71)	0.889	0.744
BioLORD-2023	0.729 (.72–.74)	0.893	0.769
PubMedBERT-base	0.781 (.77–.79)	0.945	0.821
<i>General-Domain Models</i>			
BM25	0.779 (.77–.79)	0.920	0.814
EmbeddingGemma-290m	0.834 (.83–.84)	0.964	0.866
GTE-ModernBERT-base	0.861 (.86–.87)	0.974	0.889
BGE-base-en-v1.5	0.866 (.86–.87)	0.976	0.894
<i>Our Models</i>			
SIREN-base	0.937 (.93–.94)	0.996	0.952
SIREN-base + CE	0.952 (.95–.96)	0.997	0.963

Table 2: Results on SIREN internal test set ($n \approx 9,000$ queries). All models are base size (~ 110 – 150 M params) except EmbeddingGemma (290M). SIREN outperforms its base model (+7.6pp) and MedCPT (+24.0pp). 95% bootstrap CIs in parentheses.

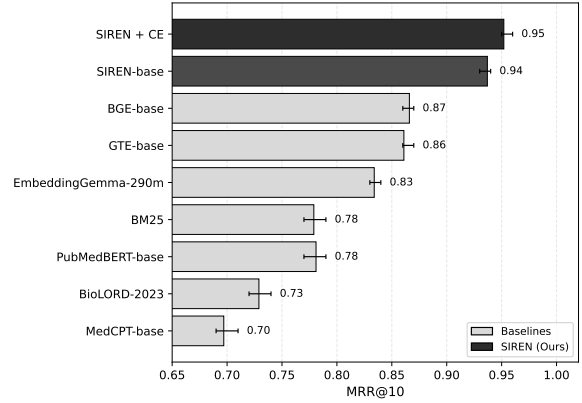


Figure 2: MRR@10 comparison on SIREN internal benchmark with 95% bootstrap confidence intervals. SIREN achieves +24pp over MedCPT.

being trained for biomedical retrieval on 255 million PubMed query-article pairs. This suggests that training on search engine logs, which capture keyword-style queries, does not transfer well to the structured inclusion criteria typical of systematic review screening.

5.2 External Benchmarks

To assess generalization, we evaluate on established benchmarks: CLEF-TAR 2019 (the gold standard for systematic review screening), CSMeD (Kusa et al., 2023) (a meta-dataset of 325 Cochrane reviews), and three biomedical tasks from BEIR (SciFact, TRECCOVID, NFCorpus). Table 3 compares SIREN against GTE-ModernBERT-base (our base model) and MedCPT.

CLEF-TAR 2019. On the standard systematic review benchmark, all models achieve near-ceiling

²<https://huggingface.co/NeuML/pubmedbert-base-embeddings>

Benchmark	Metric	SIREN	GTE	MedCPT
<i>Systematic Review</i>				
CLEF-TAR	WSS@95	0.931	0.931	0.931
CLEF-TAR	NDCG@10	0.434	0.407	0.314
CSMeD	WSS@95	0.189	0.201	0.209
CSMeD	NDCG@10	0.670	0.667	0.611
<i>BEIR Biomedical</i>				
SciFact	NDCG@10	0.770	0.753	0.708
TRECCOVID	NDCG@10	0.674	0.655	0.567
NFCorpus	NDCG@10	0.351	0.349	0.322

Table 3: External benchmark results (bi-encoder only). SIREN leads in 5 of 7 comparisons against GTE-ModernBERT-base (GTE) and MedCPT, and ties a sixth. SLERP merging preserves out-of-domain generalization.

WSS@95 (~ 0.93).³ This clustering reflects the metric’s limited discriminative power: with only 20 topics, 2.3% average relevance rate, and a coarse 95% recall threshold, all capable models achieve similar scores. More discriminative metrics reveal meaningful differences: SIREN improves NDCG@10 by +12.0pp over MedCPT (0.434 vs. 0.314). In practice, better ranking means relevant documents appear earlier, reducing the number of abstracts reviewers must read before finding useful studies.

BEIR Biomedical. SIREN outperforms both baselines on all three BEIR tasks: SciFact (+6.2pp over MedCPT), TRECCOVID (+10.7pp), and NFCorpus (+2.9pp). SIREN also exceeds its base model GTE-ModernBERT-base on all three tasks, demonstrating that our fine-tuning combined with SLERP merging preserves out-of-domain generalization.

CSMeD. SIREN underperforms MedCPT on CSMeD WSS@95 (0.189 vs. 0.209). CSMeD uses full systematic review abstracts ($\sim 1,300$ tokens) as queries, while SIREN is optimized for criteria-based queries (50–200 tokens). This query length distribution shift affects recall-focused metrics. However, SIREN wins on NDCG@10 (0.670 vs. 0.611), indicating superior ranking quality among retrieved documents. This strength could complement recall-oriented systems in hybrid retrieval workflows.

³We evaluate retrieval ranking on pooled relevance judgments, not the full active learning TAR workflow used in official CLEF-TAR runs. The high WSS@95 reflects the concentrated relevant documents in the evaluation pool.

Design Choice	Δ MRR	Note
MNRL vs CoSENT	+11.7pp	In-batch negatives
Hard negatives	+4.8pp	BM25-mined
3-class vs binary CE	+0.9pp	Enables triage
SLERP merge ($t=0.4$)	—	+2.1pp OOD

Table 4: Key design decisions from ablation studies. Δ MRR is internal benchmark improvement; SLERP row shows OOD recovery on external benchmarks. Loss function and hard negative mining provide the largest gains.

5.3 Ablations

We conducted systematic ablation studies during development to identify the most impactful design choices. Table 4 summarizes the key findings; full results appear in Appendix E.

The choice of loss function has the largest impact: MNRL with in-batch negatives outperforms CoSENT by 11.7pp. At the ablation batch size of 32, MNRL provides 31 implicit negatives per example, creating a richer contrastive signal than explicit triplet-based losses. Hard negative mining via BM25 adds 4.8pp by exposing the model to lexically similar but semantically irrelevant documents during training.

The 3-class cross-encoder slightly outperforms binary classification (+0.9pp), though confidence intervals overlap. We frame this as a design choice enabling triage workflows rather than a performance contribution: reviewers can prioritize “Relevant” documents while deferring “Partial” matches to a second pass.

Cross-encoder Transfer. Our cross-encoder helps weaker bi-encoders substantially: MedCPT + SIREN CE achieves 0.931 MRR@10 (+23pp over MedCPT alone), demonstrating that the cross-encoder captures complementary relevance signals even when paired with a different retriever.

Reranker Specialization. Conversely, pairing SIREN bi-encoder with non-screening cross-encoders degrades performance: MedCPT-CE yields 0.848 MRR@10 (−8.9pp) and BGE-reranker-v2 yields 0.846 (−9.2pp). Both biomedical and general-purpose rerankers hurt SIREN’s rankings, highlighting the importance of task-specific reranking for screening queries.

5.4 Query Type Analysis

Systematic review screening involves two distinct query types: queries seeking documents that match

Model	Pos.	Part.	Gap
<i>Biomedical Models</i>			
BioLORD-2023	0.870	0.588	28.2
PubMedBERT-base	0.916	0.646	27.0
MedCPT	0.819	0.576	24.3
<i>General-Domain Models</i>			
BM25	0.916	0.642	27.4
EmbeddingGemma-290m	0.952	0.717	23.5
GTE-ModernBERT-base	0.961	0.761	20.0
BGE-base-en-v1.5	0.957	0.776	18.1
SIREN	0.987	0.887	10.0

Table 5: MRR@10 by query type. Gap = performance drop on partial queries (pp). Biomedical models exhibit *larger* gaps (24–28pp) than general-domain dense retrievers (18–24pp). Even EmbeddingGemma (290M params) shows a 23.5pp gap. SIREN’s graded relevance training reduces this to 10.0pp.

all eligibility criteria (positive queries) and queries seeking documents that match *some but not all* criteria (partial queries). The latter is particularly challenging. A document about “RCTs of aspirin in diabetic patients” is partially relevant to a query seeking “RCTs of aspirin for cardiovascular prevention in diabetic patients over 65.” Table 5 analyzes model performance across these query types.

Biomedical-specific models show larger positive-partial gaps than general-domain models. BioLORD (28.2pp), PubMedBERT (27.0pp), and MedCPT (24.3pp) all struggle more with partial relevance than BGE-base (18.1pp) and GTE-base (20.0pp). EmbeddingGemma-290m (23.5pp gap) performs comparably to the biomedical models despite having roughly twice as many parameters, reinforcing that model scale alone does not address graded relevance. This pattern suggests that the challenge of systematic review screening is not domain vocabulary or model capacity but the training signal: models trained on binary relevance or keyword-style queries lack the graded signal needed to distinguish “fully matches criteria” from “partially matches criteria.”

SIREN’s explicit 3-class training directly addresses this gap. By training on queries that explicitly request partial matches, the model learns to represent the nuanced distinctions between full and partial relevance. The result is a 45% reduction in the positive-partial gap (10.0pp vs. 18.1–28.2pp for baselines) while simultaneously achieving the highest absolute performance on both query types.

6 Discussion

Why Ranking Quality Matters. While WSS@95 is the standard metric for systematic review screening, our results highlight the importance of ranking metrics (NDCG, MRR). On CLEF-TAR 2019, all models achieve near-identical WSS@95 (~ 0.931), yet SIREN shows +12.0pp higher NDCG@10. In practice, this means relevant documents appear earlier in the ranked list, reducing the number of abstracts reviewers must examine before finding useful studies. For a review screening 10,000 abstracts, better ranking can save hours of reviewer time even at equivalent recall.

The Training Signal Hypothesis. Our internal benchmark results suggest that the challenge of criteria-based screening is not biomedical domain knowledge (MedCPT, PubMedBERT, and BioLORD all have that) but the training signal. Models trained on binary relevance or keyword-style search queries lack the graded signal needed to distinguish full from partial criteria matches. We note that architectural differences also contribute: GTE-ModernBERT-base uses a more recent architecture than MedCPT’s BERT (Devlin et al., 2019) backbone, and the gap between biomedical and general-domain models partly reflects this. However, the +7.6pp improvement of SIREN over its own base model (GTE) isolates the effect of our training data. This finding may generalize to other specialized retrieval tasks where task-specific synthetic data could prove more effective than domain-specific pretraining.

Synthetic Data and Model Merging. Our results demonstrate that LLM-generated training data can be effective for specialized retrieval tasks, even without human annotation. The seven diversity dimensions play a key role: without them, the teacher LLM tends to generate homogeneous queries that overfit to a narrow style. SLERP merging addresses a practical challenge in fine-tuning: the fine-tuned bi-encoder (before merging) gains +5pp on the internal benchmark but loses performance on external benchmarks. By interpolating between fine-tuned and base weights ($t = 0.4$), we recover out-of-domain performance while retaining task-specific gains, consistent with the observation that fine-tuned and base models occupy nearby regions of weight space (Wortsman et al., 2022b).

Qualitative Observations. Manual inspection of system outputs across query types confirmed that SIREN correctly prioritizes documents matching all stated criteria. Common failure modes involve queries with implicit exclusion criteria not explicitly stated in the eligibility requirements, and edge cases where partial matches share nearly all criteria with fully relevant documents. These observations suggest that further gains may require richer representations of negated or implicit criteria.

Future Work. Comparing SIREN’s pipeline against direct LLM-based screening (using the teacher model at inference time) would quantify the efficiency-quality trade-off of our distillation approach. Integration with active learning frameworks like ASReview could enable iterative improvement during screening. Extension to full-text screening (beyond abstracts) would address a practical need, as some eligibility criteria can only be verified in the methods or results sections. Multilingual systematic reviews represent an underexplored challenge.

7 Conclusion

We presented SIREN, a neural systematic review screening system that improves ranking quality by +12.0pp NDCG@10 over MedCPT on CLEF-TAR 2019 and outperforms baselines on 5 of 7 external comparisons. On our released benchmark, SIREN achieves +24pp MRR@10 over MedCPT. Our analysis shows that biomedical-specific models exhibit *larger* performance gaps on partial-relevance queries than general-domain models, suggesting the challenge lies in training signal rather than domain knowledge.

Our synthetic data generation pipeline addresses this gap through explicit graded relevance training, reducing the positive-partial performance gap by 45% while achieving the highest absolute performance.

We release SIREN-60k, models, and evaluation code to support reproducible research in systematic review automation.⁴

⁴Dataset: <https://huggingface.co/datasets/Praise2112/siren-screening>; Models: <https://huggingface.co/Praise2112/siren-screening-biencoder>; Code: <https://github.com/praise2112/siren-screening>

Limitations

Our internal benchmark results (+24pp MRR@10) are measured on SIREN-60k, a synthetic benchmark generated by the same pipeline used for training data, and should be interpreted as a measure of model fit rather than real-world screening performance. We rely on CLEF-TAR, CSMed, and BEIR for generalization evidence, though we note that confidence intervals are not reported for external benchmarks and CLEF-TAR has only 20 topics. We did not compare against direct LLM-based screening (e.g., using the teacher model at inference time), which would help quantify the value of our distillation approach. Our evaluation is limited to abstract-level screening; full-text screening presents different challenges as eligibility criteria may require information from methods or results sections. We evaluated on English-language benchmarks only; multilingual systematic reviews remain unexplored. CSMed performance (WSS@95 = 0.189) lags behind MedCPT due to query format mismatch: CSMed uses full review abstracts (~1,300 tokens) as queries, far exceeding typical eligibility criteria. The cross-encoder is specialized for screening and should not be expected to generalize to other retrieval tasks. While we conducted manual inspection of system outputs (Section 6), formal evaluation with independent systematic review experts would provide stronger validation of practical utility and remains a priority for future work. The ablation studies were conducted on a development model (PubMedBERT-base) and insights were transferred to the final GTE-ModernBERT-base model, so ablation findings may not perfectly reflect the final model’s behavior.

Ethics Statement

Our system is designed as a screening aid: it prioritizes candidate documents, but final inclusion decisions remain with domain experts. We release our models and data under a permissive license to support open research in systematic review automation.

References

- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP*, pages 3615–3620.
- Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and

- Rodrigo Nogueira. 2022. InPars: Data augmentation for information retrieval using large language models. In *Proceedings of SIGIR*, pages 2622–2631.
- Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, 7(2):e012545.
- Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.
- Gordon V Cormack and Maura R Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of SIGIR*, pages 153–162.
- Gordon V Cormack and Maura R Grossman. 2016. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of CIKM*, pages 1039–1048.
- Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Baber, Jamie Callan, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *Proceedings of ICLR*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Luyu Gao and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP*, pages 316–321.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd edition. John Wiley & Sons.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.
- Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.
- Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview. In *CEUR Workshop Proceedings*, volume 1866.
- Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In *CEUR Workshop Proceedings*.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781.
- Jaana Kekäläinen. 2005. Grading relevance. *Information Processing & Management*, 41(5):1097–1113.
- Wojciech Kusa, Allan Hanbury, and Petr Knöth. 2023. CSMED: Bridging the dataset gap in automated citation screening for systematic literature reviews. In *NeurIPS Datasets and Benchmarks Track*.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rber, Nikhil Saunshi, Ramaswamy Kalb, Nishanth Dikkala, Sameer Singh, Sham Kakade, and Prateek Jain. 2022. Matryoshka representation learning. In *Proceedings of NeurIPS*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Xinyu Mao, Bevan Koopman, and Guido Zuccon. 2024. Dense retrieval with continuous feedback for systematic review screening prioritisation. In *Proceedings of SIGIR*.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.
- Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1):1–10.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.
- François Remy, Kris Demuynck, and Thomas Demeester. 2024. [BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights](#). *Journal of the American Medical Informatics Association*, page ocae029.

Eero Sormunen. 2002. Liberal relevance criteria of TREC: Counting on negligible documents? In *Proceedings of SIGIR*, pages 324–330.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of NeurIPS*.

Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahber, Jan de Boer, Felix Weijdem, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, and 1 others. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, and 1 others. 2025. **Embeddinggemma: Powerful and lightweight text representations**. *Preprint*, arXiv:2509.20354.

Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau, and Thomas A Trikalinos. 2012. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In *Proceedings of IHI*, pages 819–824.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of NAACL*, pages 2063–2077.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carber, Simon Kornblith, and Ludwig Schmidt. 2022a. Model soups: averaging weights of multiple finetuned models improves accuracy without increasing inference time. In *Proceedings of ICML*, pages 23965–23998.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2022b. Robust fine-tuning of zero-shot models. In *Proceedings of CVPR*, pages 7959–7971.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Appendix

A Synthetic Data Statistics

Table 6 summarizes the SIREN-60k synthetic training dataset.

Statistic	Value
<i>Dataset Size</i>	
Total examples	59,950
Train / Val / Test	56,950 / 1,500 / 1,500
<i>Query Type Distribution</i>	
Positive queries	50.6%
Partial queries	49.4%
<i>Query Format Distribution</i>	
Prose	87.9%
List	7.5%
PICO structured	4.7%
<i>Query Length Distribution</i>	
Short (<40 words)	80.1%
Medium (40–120 words)	19.9%
<i>Hard Negative Sampling</i>	
Teacher-generated	70%
BM25-mined	20%
Random	10%

Table 6: Statistics for the SIREN-60k synthetic training dataset.

B Query Generation Dimensions

Table 7 shows the seven diversity dimensions used in synthetic query generation.

Dimension	Values
Length	short (36%), medium (45%), long (18%)
Format	prose (29%), list (18%), structured (12%), pico (6%), logical (12%), conversational (18%)
Specificity	broad (20%), moderate (50%), narrow (30%)
Criteria count	single (20%), few (50%), many (30%)
Terminology	formal, informal, mixed, abbreviated, coded
Negation	none (50%), simple (30%), complex (10%)
Ambiguity	none (55%), moderate (27%), high (18%)

Table 7: Query generation dimensions with sampling weights (not actual distribution). Incompatible combinations are automatically resampled, so actual distribution differs (e.g., 88% prose vs. 29% weight).

C Cross-encoder Transfer Results

Table 8 shows that our cross-encoder transfers effectively to different retrievers.

D PPG Metrics

We introduce the Positive-Partial Gap (PPG) metric to measure graded relevance separation. The win

Bi-enc.	Cross-enc.	MRR	Δ
SIREN	—	0.937	—
SIREN	SIREN-CE	0.952	+1.5pp
SIREN	MedCPT-CE	0.848	−8.9pp
SIREN	BGE-reranker-v2	0.846	−9.2pp
MedCPT	—	0.697	—
MedCPT	MedCPT-CE	0.826	+12.9pp
MedCPT	SIREN-CE	0.931	+23.4pp

Table 8: Cross-encoder transfer. SIREN-CE boosts MedCPT by +23pp. Non-screening cross-encoders hurt SIREN: both MedCPT-CE (−8.9pp) and BGE-reranker (−9.2pp) degrade its rankings.

rate at margin m is the fraction of queries where the positive document scores higher than the partial document by at least m .

Model	WR@0.0	WR@0.05	Gap
MedCPT	79.2%	47.3%	0.049
SIREN	90.2%	76.1%	0.115
MedCPT + CE	96.7%	39.4%	− [†]
SIREN + CE	98.4%	97.9%	− [†]

Table 9: Positive-Partial Gap metrics. $WR@m$ = queries where positive scores higher than partial by margin m . [†]Gap undefined for CE (class probabilities).

E Training Details

All training was conducted on a single NVIDIA RTX 4090 GPU.

Phase	Experiment	MRR@10
<i>Phase 1: Loss Functions</i>		
	MNRL (winner)	0.894
	Triplet	0.833
	CoSENT	0.777
<i>Phase 2: Hard Negatives</i>		
	With BM25 negatives	0.878
	Without hard negatives	0.830
<i>Phase 3: Training Techniques</i>		
	Matryoshka (MRL)	0.883
	Layer-wise LR	0.884

Table 10: Bi-encoder ablations on validation set. Ablations on PubMedBERT-base; insights applied to final model.

Hyperparameter	Bi-enc.	Cross-enc.
Base model	GTE-MB	GTE-reranker
Optimizer	AdamW	AdamW
Learning rate	2×10^{-5}	2×10^{-5}
LR schedule	Cosine	Cosine
Batch size	512	32
Epochs	1	1
Max seq length	768	768
Warmup ratio	0.1	0.1
Weight decay	0.01	0.01
SLERP t	0.4	0.2

Table 11: Training hyperparameters. GTE-MB = GTE-ModernBERT-base.