

Modular Expert Merging for Biomedical Retrieval

Sameh Khattab^{1,4*}, Jean-Philippe Corbeil^{2*}, Osman Alperen Çinar-Koraş^{1,4},
Amin Dada¹, Julian Friedrich¹, Jiawei He³, Douglas Teodoro³, Jens Kleesiek^{1,4†}

¹Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Germany

²Microsoft Healthcare & Life Sciences, Canada

³Department of Radiology and Medical Informatics, University of Geneva, Switzerland

⁴Faculty of Computer Science, University of Duisburg-Essen, Germany

Abstract

Adapting general-purpose LLMs into domain-specialized dense retrievers typically requires large-scale training on mixed-domain data. We show that merging independently trained domain-specialized experts consistently exceeds this approach across four decoder-only LLM families (0.6B–7B), four merging methods, and twelve medical and general retrieval tasks from MTEB, suggesting that parameter-space composition captures complementary domain strengths that large-scale mixed-domain training averages out. To further maximize expert quality, we introduce Synthesize-Train-Merge (STM), a modular framework that synthesizes hard negatives with a top-tier LLM and fine-tunes domain-specialized experts via LoRA before merging them, without continual pre-training. Synthesized hard negatives yield the largest gains for smaller models, and STM achieves strong performance on biomedical retrieval tasks while maintaining competitive general-domain results across all four backbone families.

1 Introduction

Retrieval-augmented generation (RAG) (Lewis et al., 2020) has become a standard approach for grounding Large Language Models (LLMs), improving on recent knowledge and mitigating hallucinations (Xiong et al., 2024; Fan et al., 2024; Abo El-Enen et al., 2025; Ayala and Bechard, 2024; Barry et al., 2025). At the core of RAG systems are lexical and semantic retrievers (Sawarkar et al., 2024; Wang et al., 2024b), with dense retrievers built on decoder-only LLMs now achieving state-of-the-art performance on embedding tasks (Wang et al., 2024a; BehnamGhader et al., 2024). This

*Corresponding authors: sameh.khattab@uk-essen.de, jcorbeil@microsoft.com

†Other affiliations: Cancer Research Center Cologne Essen (CCCE), German Cancer Consortium (DKTK, Partner site Essen) and Department of Physics of TU Dortmund, Germany.

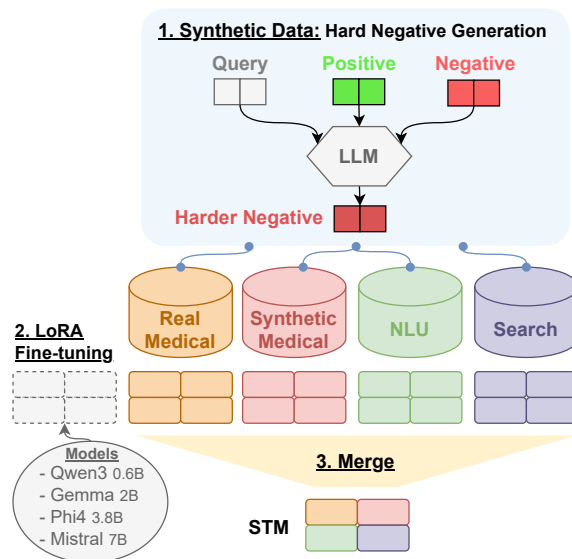


Figure 1: Diagram of our recipe to obtain the *STM* retrievers: 1) synthetic data based on hard negative generation, 2) LoRA fine-tuning, and 3) model merging. We segment the BMRetriever dataset into four splits: *Real Medical*, *Synthetic Medical*, *NLU*, and *Search*.

suggests that general-purpose LLMs already provide a strong foundation for retrieval.

Yet adapting these models to domain-specific retrieval, such as in the biomedical field, remains an open challenge. Prior work (Xu et al., 2024) shows that contrastive pre-training on large corpora followed by instruction tuning yields strong medical retrievers, and hard negative mining further boosts performance (de Souza Pereira Moreira et al., 2024; Lee et al., 2024; Shao et al., 2025). However, the computational cost of continual pre-training is substantial, and it remains unclear which subsets of the data mixture drive most of the gains, or whether top-tier LLMs can be leveraged to synthesize effective hard negatives in place of mined ones.

In parallel, model merging (Goddard et al., 2024) has emerged as a powerful paradigm for composing robust models (Wortsman et al., 2022; Aakanksha

et al., 2024), enabling modular development (Corbeil et al., 2025) and efficient data ablation (Na et al., 2024). Basic techniques such as ModelSoup (Wortsman et al., 2022) are now part of the training recipes of several recent retrievers, including Granite (Awasthy et al., 2025), Llama-Embed-Nemotron-8B (Babakhin et al., 2025), EmbeddingGemma (Vera et al., 2025), and Qwen3 Embedding (Zhang et al., 2025). To our knowledge, only Xiao et al. (2024) has systematically studied ModelSoup merging (i.e., linear merging) for both encoder and decoder retrievers. Beyond this, little is known about whether merging gains hold consistently across model scales and data mixtures, which data subsets matter most, and how to best optimize merging coefficients for retrieval.

We ask whether merging domain-specialized experts can replace large-scale mixed-domain training entirely. Across four decoder-only LLM families and twelve retrieval tasks, we find consistently that it can. To maximize expert quality before merging, we introduce **Synthesize-Train-Merge (STM)**, a modular framework that synthesizes hard negatives with an LLM, fine-tunes domain-specialized experts via LoRA, and merges them in parameter space; all without continual pre-training (Figure 1). Our contributions are as follows:

- We demonstrate that parameter-space **merging of dense retrieval experts consistently performs above large-scale mixed-domain fine-tuning** across four decoder-only LLM families (0.6B–7B), four merging methods, and twelve retrieval tasks from MTEB (Figure 2).
- We introduce **STM (Synthesize-Train-Merge)**, a modular framework for building domain-specialized dense retrievers without continual pre-training.
- We show the practical effectiveness of **LLM-synthesized hard negatives** for retrieval expert training across a range of backbone families, with especially strong benefits for smaller models.
- We release **code**¹, an improved **retrieval dataset**, and **STM model checkpoints**² to facilitate reproducibility and future research.

¹<https://github.com/TIO-IKIM/STM>

²<https://huggingface.co/collections/ikim-uk-essen/stm>.

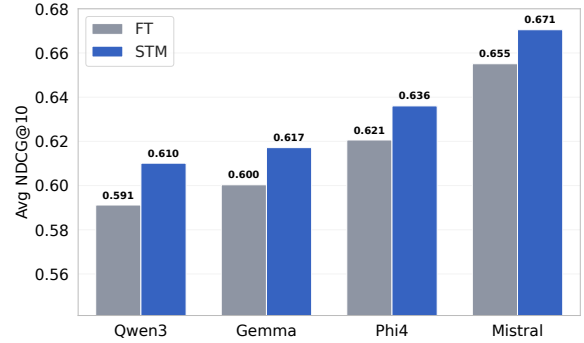


Figure 2: Performance comparison of STM Merged Models versus models fine-tuned on the combined datasets of all merged experts, across 4 base models, using the average NDCG@10 metric across 12 MTEB tasks.

	LLM2Vec	BMRetriever	STM (Ours)
Pooling	Average	EOS	EOS
Training Setup	LoRA	LoRA	LoRA
Training Recipe	MNTP + SimSCE	PT + FT	FT + Merging
Negatives	SimSCE	Sampled Top-k	Synthetic
Dataset size	1.5M	11.4M	1.4M
Domain	General	Biomedical	General & Biomedical

Table 1: Comparison of attributes between previous methods (LLM2Vec, BMRetriever) and ours. LLM2Vec is a multi-task contrastive embedding model, and BMRetriever a biomedical dense retriever pretrained with pseudo-queries (PT) and finetuned on a mix of real and synthetic labeled data with mined hard negatives (FT).

2 Related Works

2.1 Retrievers from Decoder-Only Models

E5 (Wang et al., 2024a) first demonstrated state-of-the-art performance on the MTEB benchmark (Muennighoff et al., 2022) from pretraining and fine-tuning Mistral 7B (Jiang et al., 2023), a decoder-only model, on real and synthetic paired retrieval data. LLM2Vec (BehnamGhader et al., 2024) showed that using bidirectional attention masking with average pooling on LLMs, and training them with *masked next token prediction* and SimCSE (Gao et al., 2021) led to improvements on retrieval tasks. NV-Embed (Lee et al., 2024) introduced the *latent attention* pooling layer during training with positive-aware hard negatives and non-retrieval task data to reach stronger performance. BMRetriever (Xu et al., 2024) lever-

aged a vast contrastive pre-training, followed by instruction tuning on a mix of real and synthetic datasets, yielding strong dense retrieval models for the biomedical domain.

2.2 Hard Negative Mining

Hard negative mining has become a crucial design choice for training modern dense retrievers. Classical negative mining schemes (Xiong et al., 2021; Zhan et al., 2021) showed that retrieving top-ranked passages during training can accelerate contrastive learning. Despite its limitation to exact word overlap, BM25 (Robertson and Zaragoza, 2009) is still widely used to surface hard negatives (Karpukhin et al., 2020; Zhan et al., 2021). NV-Retriever (de Souza Pereira Moreira et al., 2024) revisits this space with *positive-aware* hard negative mining, which boosts the performance on the MTEB benchmark.

Beyond mined hard negatives, a small but growing line of work starts to explore *generated* hard negatives using LLMs. SyNeg (Li et al., 2024) uses an LLM with a multi-attribute self-reflection prompting strategy to synthesize hard negatives and combines them with retrieved negatives in a hybrid sampling scheme. Their ablation study shows that gains only arise from the hybrid method. In contrast, we show that only prompting a top-tier LLM to generate hard negatives can yield substantial gains, particularly for smaller models.

2.3 Model Merging

Linear-mode connectivity (Frankle et al., 2020; Mirzadeh et al., 2020) established that independently trained solutions can be connected by low-loss paths in parameter space, motivating model combination methods based on interpolation. Building on this insight, Model Soup (Wortsman et al., 2022) showed that averaging multiple training checkpoints can outperform selecting a single one. These ideas naturally extend from combining checkpoints of a single model to merging distinct expert models: *task arithmetic* (Ilharco et al., 2022) formulates merging as adding and subtracting task-specific deltas from a base model, while methods such as Ties-merging (Yadav et al., 2023) and DARE (Yu et al., 2024) explicitly tackle parameter interference when merging multiple experts. Recent work further shows that such parameter-level merging can be competitive with data-mixing strategies (Aakanksha et al., 2024; Na et al., 2024).

Authors (Labrak et al., 2024; Corbeil et al., 2025)

employed merging to build robust medical LLMs. EmbeddingGemma (Vera et al., 2025) and Qwen3 Embedding (Zhang et al., 2025) exploit merging in their recipe without studying its impact.

3 Methods

3.1 Hard Negative Generation

Training dense retrievers with contrastive objectives requires negatives that are both topically related and semantically distinct from the positives. Existing hard negative mining strategies frequently struggle to balance informativeness and correctness, often introducing either trivial negatives or false negatives that are actually relevant (de Souza Pereira Moreira et al., 2024).

To alleviate this issue, we employ GPT-5 (Singh et al., 2026) to generate synthetic hard negatives. Given a query q , a positive passage p^+ , and an existing mined negative p^- , we prompt the LLM to generate a new negative passage $\tilde{p}^-$ that remains lexically and topically aligned with q while being semantically irrelevant or contradictory. The prompt provides the full context (q, p^+, p^-) and explicitly instructs the model to preserve surface-level similarity while altering semantic intent. The exact prompt template is provided in Appendix B.1.

3.2 Instruction Fine-Tuning

We fine-tune decoder-only backbone models using a contrastive learning objective to obtain dense retrievers. We adopt the InfoNCE loss (van den Oord et al., 2018), which encourages each query embedding to be closer to its corresponding positive passage than to all other passages in the batch and to any provided hard negative passage.

Formally, given a batch of N triplets $\{(q_i, p_i^+, p_i^-)\}_{i=1}^N$, the loss term for a given query q_i is defined as:

$$\mathcal{L}_i = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^-)} + \sum_{j=1}^N e^{\text{sim}(q_i, p_j^+)}}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity between embeddings, and the denominator includes two terms: the first one uses a hard negative p_i^- , and the second one employs other passages p_j^+ for $j \neq i$ as in-batch negatives. We average over the loss terms in the batch to obtain the total loss.

3.2.1 Expert Model Fine-Tuning

To enable modular specialization, we fine-tune multiple expert retrievers on different, coherent subsets of the BMRetriever fine-tuning dataset (Xu et al., 2024): *medical synthetic*, *medical real*, *natural language understanding (NLU)*, and *search*.

The *medical real* subset includes sentence-level biomedical inference and similarity datasets such as MedNLI (Shivade, 2017) and Medical Question Pairs (McCreery et al., 2020), as well as passage-level biomedical QA benchmarks including MEDIQA (Ben Abacha et al., 2019), medical StackExchange QA (Team, 2021), and medical dialogue data (Li et al., 2023). The *medical synthetic* subset consists of LLM-generated biomedical retrieval pairs. To improve general-domain relevance modeling, the dataset further incorporates *NLU* benchmarks such as Natural Questions (Kwiatkowski et al., 2019), FEVER (Thorne et al., 2018), ELI5 (Fan et al., 2019), SNLI (Bowman et al., 2015), and the MS MARCO passage ranking dataset (Bajaj et al., 2018) as the *search* subset.

We train four experts per backbone model, each emphasizing a distinct data composition or training configuration, such as synthetic hard negatives.

3.3 Model Merging

Model merging aims to combine multiple expert retrievers into a single model that inherits complementary strengths without additional training. Given a set of expert models $\{M_k\}_{k=1}^K$ with parameters $\{\theta_k\}$, merging methods compute a unified model $\hat{M}$ by operating directly in parameter space.

In the simplest case, linear merging (Wortsman et al., 2022) is defined as:

$$\theta_{\hat{M}} = \sum_{k=1}^K \alpha_k \theta_k, \quad (1)$$

where α_k are the *weight* coefficients with $0 \leq \alpha_k \leq 1$.

The task arithmetic approach (Ilharco et al., 2022) relies instead on a linear combination of task vectors $\tau_k = \theta_k - \theta_B$, which is the delta of parameters between the k^{th} expert and the parameters of the base model θ_B . The merged model becomes

$$\theta_{\hat{M}} = \theta_B + \sum_{k=1}^K \alpha_k \tau_k, \quad (2)$$

Ties merging (Yadav et al., 2023) also leverages task vectors τ_k with two strategies to mitigate the *task-interference* phenomenon: keeping only high-magnitude parameter changes by introducing a second parameter named density $\delta_k \in [0, 1]$, and the sign agreement algorithm, which is a majority vote on the signs across all τ_k . The DARE-Ties technique (Yu et al., 2024) replaces the magnitude-based pruning of Ties with random dropout at rate δ_k followed by a rescaling of task vectors.

4 Experiments

4.1 Finetuning Expert Datasets

For fine-tuning, we reuse the BMRetriever data mixture (Xu et al., 2024) which originally contained 14 data sources. We reorganize it into the four coherent subsets in Table 2 balanced by size, task, and domain.

	Medical		General	
Split	Med-Synth	Med-Real	Search	NLU
#Pairs	431,000	306,000	438,000	251,000

Table 2: Pair counts for our custom splits of the BMRetriever fine-tuning dataset, comprising four splits: two in the medical domain and two in the general domain.

4.2 Training Setup

We experiment with four decoder-only backbone architectures of varying sizes; Qwen3 (Yang et al., 2025), Gemma (Team et al., 2024), Phi-4 (Abdin et al., 2024), and Mistral (Jiang et al., 2023); the configurations are summarized in Table 3.

	Qwen3	Gemma	Phi4	Mistral
Parameters	0.6B	2B	3.8B	7B
Dimensions	1,024	2,048	3,072	4,096

Table 3: Backbone model used for expert training. *Parameters* indicate the total model size, and *Dimensions* refers to the hidden size.

In preliminary experiments, we consider both EOS-token pooling and mean pooling strategies. However, we observe that the former yields slightly stronger performance. Therefore, we adopt EOS pooling for all reported results. Table 1 compares our implementations with previous works.

During fine-tuning, retrieval prompts are prepended to each query, and models are trained using the InfoNCE loss. We fine-tune all models using LoRA adapters (Hu et al., 2021) applied to all

linear layers following prior work (BehnamGhader et al., 2024; Lee et al., 2024; Xu et al., 2024). Hyperparameters, including learning rate, batch size, number of steps, and LoRA configuration, are summarized in Table 9 of Appendix A.

We conduct fine-tuning under several configurations. We begin with standard fine-tuning on the base dataset and the four data splits. We then fine-tune the models on the datasets augmented with synthetic hard negatives. Finally, we evaluate a configuration that combines the best experts from either standard fine-tuning or synthetic hard-negative variants.

4.3 Model Merging

We perform model merging using MergeKit (Goddard et al., 2024), which supports parameter-space merging for HuggingFace-compatible models. We evaluate four main merging strategies: linear interpolation (Wortsman et al., 2022), task arithmetic (Ilharco et al., 2022), Ties merging (Yadav et al., 2023), and DARE-Ties (Yu et al., 2024).

We compare the selection of the best merged models across two different coefficient-optimization techniques: (1) a *random* search approach of N sampling rounds, and (2) the CMA-ES *evolutionary* algorithm (Akiba et al., 2025) at N evaluations based on the MergeKit implementation³. For linear merging and task arithmetic, we sweep weight coefficients $\alpha \in \{0, 0.1, \dots, 0.9\}$. For Ties and DARE-Ties, we sweep the weight coefficients over the same range, and vary the density parameter over $\rho \in \{0.1, 0.2, \dots, 0.9\}$.

Without any training, all merged models are evaluated on a frozen development split comprising four retrieval tasks: NFCorpus (Boteva et al., 2016), FiQA-2018 (Thakur et al., 2021), FEVER (Thorne et al., 2018), and CLUE-IR (Corbeil et al., 2025), spanning biomedical in-domain and general out-of-domain retrieval. For NFCorpus and FiQA-2018, dev queries are sampled from the official validation split and are disjoint from the full test corpora used in MTEB evaluation; FEVER and CLUE-IR are held exclusively for tuning and do not appear in the benchmark. Hyperparameters are selected using unweighted mean NDCG@10 across the four subsets. Full details of the frozen development benchmark construction, sampling procedure, search configuration, and merge-selection protocol are provided in Appendix C.1 and Appendix C.3.

³Adapted for the evaluation of retrieval tasks.

4.4 Baselines

We compare our models against a diverse set of retrieval baselines, including BM25 (Lù, 2024), Contriever (Izacard et al., 2021), E5-v2 (Wang et al., 2024a), GTR (Ni et al., 2021), LLM2Vec 3B (BehnamGhader et al., 2024), BMRetriever 2B (Xu et al., 2024), and Promptriever 7B (Weller et al., 2024). Baselines are selected to match our models in parameter scale or architectural family and to reflect comparable training budgets.

4.5 Evaluation

Previous works have focused on either the medical domain or the general domain. In this work, we aim at maximizing medical performance while maintaining general retrieval capabilities. We evaluate retrieval performance on the English medical subset of the Medical MTEB benchmark (Muennighoff et al., 2022; Enevoldsen et al., 2025), which includes TREC-COVID (Roberts et al., 2021), SciFact (Wadden et al., 2020), NFCorpus (Boteva et al., 2016), CUREv1 (Athar Sheikh et al., 2025), PublicHealthQA (Xing Han Lu, 2024), FeedbackQA (Li et al., 2022) and MedicalQA (Ben Abacha and Demner-Fushman, 2019). To assess general-domain generalization, we additionally evaluate on five general-domain MTEB datasets, including FiQA (Thakur et al., 2021), ArguAna (Wachsmuth et al., 2018), SciDocs (Cohan et al., 2020), and two NanoMTEB subsets (NanoFEVER (Thorne et al., 2018) and NanoQuora (DataCanary et al., 2017)).

Following prior work on dense retriever evaluation (Xu et al., 2024; Weller et al., 2024; Muennighoff et al., 2022), we report NDCG@10 and Recall@100 on established public benchmarks. Scores are averaged across three evaluation groups: medical datasets, general-domain datasets, and the full dataset collection. For instruction-tuned retrievers, we use a shared retrieval prompt across all models at evaluation time.

Descriptions of the twelve tasks, including corpus sizes and query types, are in Appendix C.2.

5 Results and Analysis

5.1 Comparison with Prior Retrievers

We evaluate the best-merged STM models against established baselines on twelve MTEB tasks (Table 4), following established practice in retrieval evaluation (Muennighoff et al., 2022; Enevoldsen et al., 2025; Xu et al., 2024).

Model	Size	Avg Med	Avg General	Avg All
BM25	-	0.532	0.515	0.525
Contriever	150M	0.508	0.533	0.519
E5 Large V2	335M	0.654	0.576	0.622
GTR T5 XL	1.2B	0.581	0.586	0.583
BMRetriever	2B	0.645	0.560	0.609
LLM2Vec	3B	0.635	0.597	0.619
Promptriever	7B	0.682	0.605	0.650
STM_{Qwen3}	0.6B	0.636	0.574	0.610
STM_{Gemma}	2B	0.648	0.573	0.617
STM_{Phi4}	3.8B	0.659	0.604	0.636
$STM_{Mistral}$	7B	0.693	0.639	0.671

Table 4: Summary of retrieval performance against baselines across medical and general domains. NDCG@10 scores are averaged across 12 MTEB tasks. Bold indicates the highest value per column. Full per-task results are reported in Appendix C, Table 12.

$STM_{Mistral}$ records the highest mean NDCG@10 across the twelve tasks (0.671), 0.021 above the strongest baseline, Promptriever 7B (0.650), and 0.049 above E5 Large V2 (0.622). BMRetriever shows higher medical than general scores (0.645/0.560), indicating biomedical specialization at the cost of general-domain coverage. STM models show a more balanced profile across both subsets. STM_{Gemma} records a higher general score than BMRetriever (0.573 vs. 0.560) despite sharing the same 2B base model, without sacrificing medical performance (0.648 vs. 0.645). STM_{Qwen3} (0.6B) records higher mean NDCG@10 than GTR T5 XL (1.2B) (0.610 vs. 0.583). Recall@100 results are consistent with these findings (Table 10).

5.2 Model Merging

Table 5 shows that merged models record higher mean NDCG@10 than fully fine-tuned counterparts across all backbone families and model sizes (Figure 2). TIES and Task Arithmetic achieve the same average NDCG@10 across the four families (0.632), with TIES leading for Mistral (0.671), Task Arithmetic tying with TIES for Gemma (0.617), and DARE-TIES leading for Phi4 (0.636). This observation suggests mixed results with respect to merging techniques; while simple weighted arithmetic on task vectors is sufficient in certain settings, more complex interference-aware methods offer an advantage in others. Interference-aware methods (TIES, DARE-TIES) are preferable when the risk

Method	Qwen3	Gemma	Phi4	Mistral
Full Dataset Fine-Tuning				
FT-all _{HNM}	0.591	0.600	0.621	0.655
FT-all _{SHN}	0.600	0.605	0.629	0.648
Modular Specialized Experts				
Best Expert _{HNM}	0.583	0.591	0.611	0.652
Best Expert _{SHN}	0.592	0.598	0.605	0.647
Merged Experts (STM)				
STM-DARE	0.608	0.615	0.636	0.660
STM-TIES	0.608	0.617	0.630	0.671
STM-Linear	0.610	0.608	0.623	0.662
STM-TA	0.607	0.617	0.634	0.668

Table 5: Comparison of full-dataset fine-tuning, modular specialized experts, and STM-based expert merging across four backbone families. NDCG@10 scores are averaged across 12 MTEB tasks. Bold indicates the best-performing result within each backbone family. Full results are reported in Appendix C, Table 11.

of task interference is non-negligible, consistent with the high sign-conflict rate we measure across expert task vectors (~87% across all four backbones; Table 20), which motivates their use as the number of experts or interference grows.

In all four families, the best STM variant records higher mean NDCG@10 than both the strongest individual expert and full-dataset fine-tuning, confirming that merging captures complementary domain strengths rather than amplifying a single dominant expert.

Figure 6 in Appendix C offers a mechanistic perspective on the observed benchmark advantage of merging over full-data fine-tuning. Full-data fine-tuning (dashed gray) produces larger per-layer weight deviations from the base model than any individual expert, suggesting it overwrites rather than preserves domain-specific structure. The best merged model (solid red) instead stays closer to its expert parents in parameter space, inheriting their complementary specializations without the destructive interference that joint training introduces. This pattern is consistent across all four backbone families.

5.3 Synthetic Hard Negatives

5.3.1 Expert-level Ablations

In Table 6, the synthetic hard negative approach (SHN) yields consistent improvements over standard hard-negative mining (HNM) for three of the four backbones, with average NDCG@10 gains of +4.2%, +2.0%, and +2.4% for Qwen3-0.6B,

Model	Med-Synth	Med-Real	Search	NLU	Avg
Qwen3 0.6B					
FT _{HNM}	0.583	0.568	0.506	0.546	0.551
FT _{SHN}	0.592	0.583	0.557	0.563	0.574
$\Delta\%$	+1.5%	+2.6%	+10.1%	+3.1%	+4.2%
Gemma 2B					
FT _{HNM}	0.581	0.591	0.503	0.520	0.549
FT _{SHN}	0.598	0.575	0.508	0.559	0.560
$\Delta\%$	+2.9%	-2.7%	+1.0%	+7.5%	+2.0%
Phi4 3.8B					
FT _{HNM}	0.611	0.611	0.497	0.585	0.576
FT _{SHN}	0.605	0.600	0.591	0.566	0.590
$\Delta\%$	-1.0%	-1.8%	+18.9%	-3.2%	+2.4%
Mistral 7B					
FT _{HNM}	0.638	0.652	0.610	0.603	0.626
FT _{SHN}	0.644	0.647	0.571	0.620	0.621
$\Delta\%$	+0.9%	-0.8%	-6.4%	+2.8%	-0.8%

Table 6: NDCG@10 scores averaged over 12 MTEB tasks for expert models trained with standard hard-negative mining (HNM) and synthetic hard negatives (SHN). $\Delta\%$ denotes the relative percentage improvement of SHN over HNM.

Gemma-2B, and Phi4-3.8B, respectively. These gains are driven mainly by the general-domain Search split (+10.1% for Qwen3, +18.9% for Phi4), while the biomedical splits show mixed or negative changes. Mistral-7B is the sole exception, showing a slight average decline of -0.8%, suggesting that larger, higher-capacity models may already learn sufficiently discriminative representations from mined hard negatives, leaving less room for synthetic augmentation to contribute.

An inverse scaling trend also emerges: smaller models derive the largest average benefit from SHN (+4.2% for Qwen3-0.6B versus -0.8% for Mistral-7B). We hypothesize that smaller encoders, having more limited representational capacity, are more sensitive to the quality of negative examples during contrastive training. Our results are consistent with recent findings on scaling laws for dense retrieval (Fang et al., 2024), which show that model size and training data quality interact predictably, though the specific interaction with synthetic hard negatives warrants further study.

We verified a stratified sample of the SHN dataset using an independent embedding model not used in STM training. GPT-5 negatives occupy the expected hard-negative regime (Table 22): their mean query–passage cosine similarity (0.56)

Pool	Qwen3	Gemma	Phi4	Mistral
base	0.607	0.614	0.625	0.667
shn	0.605	0.589	0.616	0.653
best	0.604	0.617	0.634	0.668

Table 7: Mean NDCG@10 for Task Arithmetic merging across expert pools. base: standard experts; shn: synthetic hard-negative experts; best: architecture-specific mix.

is comparable to mined negatives (0.54) and well below positives (0.73), while random negatives fall to 0.20. A two-judge false-negative audit with two LLMs distinct from the generator (Table 21) confirms 95–96% of GPT-5 negatives as topically related to the query but not correct answers, with 96% raw inter-judge agreement.

5.3.2 Ablation of Modular Expert Selection

Table 7 shows that the best pool, a curated mix of standard and SHN adapters, records the highest mean NDCG@10 for Gemma (0.617), Phi4 (0.634), and Mistral (0.668). This is particularly notable for Phi4 and Mistral, where SHN provided little or no improvement at the single-expert level (Table 6), yet SHN adapters contribute meaningfully when composed via merging. The reverse holds for Qwen3: despite benefiting most from SHN at the single-expert level, all pools perform within 0.003 points of each other at merging, consistent with the inverse scaling trend in Table 6, suggesting that at 0.6B, the base experts already capture sufficient complementarity for merging to exploit. Full results across all methods are in Table 15.

5.4 Merging Coefficients

5.4.1 Optimization of Coefficients

Table 16 compares CMA-ES evolve and random search across all four model families. The two strategies achieve comparable peak performance, and neither dominates consistently across architectures or methods. Random search with $N=10$ samples is therefore a competitive and substantially cheaper alternative. For Phi4, random search records higher mean NDCG@10 than evolve on three of four methods (e.g., DARE-TIES: 0.636 vs. 0.624 for evolve), which may reflect the higher-dimensional weight space making CMA-ES convergence harder within 50 evaluations. For Mistral, random search records higher mean NDCG@10 for Task Arithmetic (0.668 vs. 0.653) and TIES (0.671 vs. 0.665). These results suggest that practitioners can obtain near-optimal merged models with a

lightweight random sweep over a coarse coefficient grid, without requiring evolutionary optimization.

5.4.2 Analysis of Coefficients

Merge weight coefficients for the best configuration per method and backbone are depicted in Figure 5⁴ of Appendix C.3. Medical experts receive the highest weights across all families, with MedReal consistently dominant. As detailed in Table 16, several optimal configurations assign near-zero weight to at least one expert: NLU is down-weighted to 0.01 in Gemma’s best Task Arithmetic configuration (MedReal 0.84, Search 0.41), Search to 0.06 in Phi4’s Task Arithmetic base configuration (MedReal 0.43), and Search to 0.01 in Mistral’s Task Arithmetic hn configuration (MedReal 0.41, MedSynth 0.44). The redundant expert varies by architecture, confirming that the optimal subset emerges naturally from the search process rather than requiring manual selection.

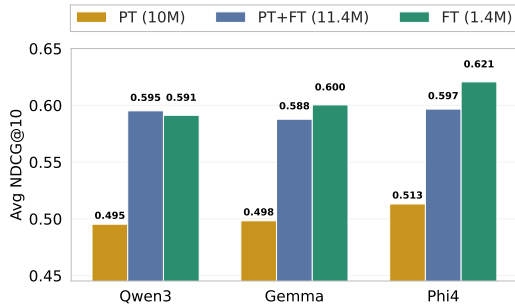


Figure 3: Performance averages of three base models pre-trained (PT) and fine-tuned (FT) on BMRetriever of 10M and 1.4M samples, respectively.

5.5 What about pretraining?

We displayed in Figure 3 results comparing pre-training (PT) and fine-tuning (FT) setups. Despite several previous works on both general (Wang et al., 2024a; Lee et al., 2024) and medical domains (Xu et al., 2024) employing an expansive pretraining phase, we observe that pretraining on 10M unlabeled pairs underperforms models fine-tuned only on 1.4M pairs. We therefore drop the pretraining step and use fine-tuning only in our experiments, reducing the training compute of a single fine-tuning run by approximately 88% relative to a PT+FT pipeline (Table 17).

⁴Density coefficients for TIES are excluded as weight coefficients directly modulate expert amplitude.

Family	FT-all	Base Merge	Merge Method
Qwen3 0.6B	0.591	0.610	Linear
Gemma 2B	0.600	0.615	DARE-TIES
Phi4 3.8B	0.621	0.636	DARE-TIES
Mistral 7B	0.655	0.667	Task-Arith

Table 8: Mean NDCG@10 across 12 MTEB tasks for merged models built exclusively from experts trained on mined negatives (Base Merge), compared against full-data fine-tuning on the same data (FT-all). Full per-task results are reported in Appendix C, Table 18.

6 Discussion

Architecture vs. scale. A confounding factor of our setup is that each scale is represented by a single architecture, which prevents disentangling the effects of model family and size. However, our main findings are consistent across all four backbone families, indicating that overall the gains from expert merging are not architecture-specific. Rather than attributing improvements to scale, we interpret these results as evidence of robust cross-architecture behavior. While we study multiple scales, evaluating several sizes within a single family would more precisely disentangle scaling effects for one fixed architecture, which we leave to future work.

On the role and cost of modularity. A natural baseline to model merging is to fine-tune a single model on the union of all data. However, our results consistently show that merging specialized experts performs above full-data fine-tuning, suggesting that joint training tends to average or overwrite domain-specific signals, while modular training preserves them and enables their effective composition. We use identical fine-tuning hyperparameters per backbone in both settings (Table 9), and as shown in Table 8, this effect is not attributable to our GPT-5 synthetic hard negatives: merging base experts without SHN still records higher mean NDCG@10 than full-data fine-tuning across all four backbones. Both settings still include the LLM-generated Med-Synth pairs from the original BMRetriever mixture, so the comparison isolates the contribution of our GPT-5 hard negatives rather than of LLM-generated data in general. This supports a view of merging not as added complexity, but as a way to better retain complementary expertise.

In practice, the added complexity is minimal: merging is performed on CPU offline on a small

development set, using simple coefficient search, and requires no additional training. The training compute for the four experts is approximately equal to that of a single full-data fine-tuning run, since the splits cover essentially the same $\sim 1.4\text{M}$ -pair dataset. The observed variation in wall-clock across backbones (Table 19) reflects implementation-level throughput rather than a systematic difference. Compared to rank-based fusion methods (Cormack et al., 2009), the key distinction is at inference time: ensembling requires running multiple models, whereas merging produces a single model, yielding a better efficiency–performance trade-off.

Why multiple experts. Finally, the use of multiple experts reflects complementary signals (real, synthetic, NLU, search), and our results indicate that combining them improves performance over both individual experts and joint training, suggesting that these signals are not redundant but synergistic.

7 Conclusion

We showed that merging domain-specialized dense retrieval experts consistently performs above large-scale mixed-domain training across four decoder-only LLM families (0.6B–7B), four merging methods, and twelve MTEB tasks, suggesting that parameter-space composition captures complementary domain strengths that mixed-domain training averages out. STM synthesizes hard negatives with a top-tier LLM, fine-tunes domain-specialized experts via LoRA, and merges them in parameter space without continual pre-training. Synthesized hard negatives yield the largest gains for smaller models, and optimal merge configurations naturally down-weight redundant experts, confirming that merging exploits complementarity rather than averaging it out. These results suggest that modular expert composition is a practical and effective alternative to mixed-domain training for building versatile dense retrievers.

Limitations

Our study has three main limitations. First, our synthetic hard negative generation relies on one proprietary LLM; we do not evaluate sensitivity to the choice of generator model, and performance may vary with open-weight alternatives, though we note that hard negative generation is a one-time

offline step that does not affect inference cost. Second, all fine-tuning and merging runs use a single random seed per configuration; multi-seed evaluation could further quantify variability. Third, our experiments are conducted exclusively on English retrieval tasks; extending STM to multilingual or cross-lingual settings is an important direction given the growing need for domain-specialized retrieval in non-English biomedical literature.

Acknowledgments

We thank François Beaulieu and Paul Vozila for their contributions to earlier versions of this work. We acknowledge the use of AI-assisted tools to improve writing clarity and assist with coding. All AI-assisted outputs were reviewed, edited, and validated by the authors. This publication was supported by the German Federal Ministry of Education and Research (BMBF) Network of University Medicine 2.0: "NUM 2.0", Grant No. 01KX2121, Project: RACOON-BI.

References

- Aakanksha, Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2024. [Mix data or merge models? optimizing for diverse multi-task learning](#). *Preprint*, arXiv:2410.10801.
- Marah I Abidin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio CT Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. [Phi-4 technical report](#). Technical Report MSR-TR-2024-57, Microsoft.
- Mohamed Abo El-Enen, Sally Saad, and Taymoor Nazmy. 2025. [A survey on retrieval-augmentation generation \(rag\) models for healthcare applications](#). *Neural Computing and Applications*, 37(33):28191–28267.
- Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2025. [Evolutionary optimization of model merging recipes](#). *Nature Machine Intelligence*, 7(2):195–204.
- Nadia Athar Sheikh, Daniel Buades Marcos, Anne-Laure Jousse, Akintunde Oladipo, Olivier Rousseau, and Jimmy Lin. 2025. [Cure: A dataset for clinical understanding & retrieval evaluation](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5270–5277. ACM.

- Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, and 3 others. 2025. [Granite embedding models](#). *Preprint*, arXiv:2502.20204.
- Orlando Ayala and Patrice Bechard. 2024. [Reducing hallucination in structured outputs via retrieval-augmented generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238, Mexico City, Mexico. Association for Computational Linguistics.
- Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. [Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks](#). *Preprint*, arXiv:2511.07025.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. [Ms marco: A human generated machine reading comprehension dataset](#). *Preprint*, arXiv:1611.09268.
- Mariam Barry, Gaetan Caillaut, Pierre Halftermeyer, Raheel Qader, Mehdi Mouayad, Fabrice Le Deit, Dimitri Cariolaro, and Joseph Gesnoui. 2025. [GraphRAG: Leveraging graph-based efficiency to minimize hallucinations in LLM-driven RAG for finance data](#). In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pages 54–65, Abu Dhabi, UAE. International Committee on Computational Linguistics.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. [Llm2vec: Large language models are secretly powerful text encoders](#). *Preprint*, arXiv:2404.05961.
- Asma Ben Abacha and Dina Demner-Fushman. 2019. [A question-entailment approach to question answering](#). *BMC Bioinformatics*, 20(1):511.
- Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. [Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.
- Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. [A full-text learning to rank dataset for medical information retrieval](#). In *Advances in Information Retrieval*, pages 716–722, Cham. Springer International Publishing.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. [SPECTER: Document-level representation learning using citation-informed transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.
- Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. 2025. [A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19352–19374, Vienna, Austria. Association for Computational Linguistics.
- Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09*, page 758–759, New York, NY, USA. Association for Computing Machinery.
- DataCanary, Lili Jiang hilfialkaff, Meg Risdal, Nikhil Dandekar, and tomtung. 2017. [Quora question pairs](#).
- Gabriel de Souza Pereira Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. [Nv-retriever: Improving text embedding models with effective hard-negative mining](#). *ArXiv*, abs/2407.15831.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). *Preprint*, arXiv:2606.19348.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatin, and 67 others. 2025. [Mmteb: Massive multilingual text embedding benchmark](#). *arXiv preprint arXiv:2502.13595*.

- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long form question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. [A survey on rag meeting llms: Towards retrieval-augmented large language models](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24*, page 6491–6501, New York, NY, USA. Association for Computing Machinery.
- Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. [Scaling laws for dense retrieval](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, page 1339–1349, New York, NY, USA. Association for Computing Machinery.
- Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning, ICML'20*. JMLR.org.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. [Arcee's MergeKit: A toolkit for merging large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.
- J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *ArXiv*, abs/2106.09685.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hananeh Hajishirzi, and Ali Farhadi. 2022. [Editing models with task arithmetic](#). *ArXiv*, abs/2212.04089.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. [Unsupervised dense information retrieval with contrastive learning](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *ArXiv*:2310.06825 [cs].
- Jared Kaplan, Sam McCandlish, Thomas Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *ArXiv*, abs/2001.08361.
- 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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [BioMistral: A collection of open-source pretrained large language models for medical domains](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. [Nv-embed: Improved techniques for training llms as generalist embedding models](#). *ArXiv*, abs/2405.17428.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen-tau Yih, Tim Rock  tschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Xiaopeng Li, Xiangyang Li, Hao Zhang, Zhaocheng Du, Pengyue Jia, Yichao Wang, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024. [Syneg: Llm-driven synthetic hard-negatives for dense retrieval](#). *Preprint*, arXiv:2412.17250.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steven Jiang, and You Zhang. 2023. [Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai \(llama\) using medical domain knowledge](#). *Cureus*, 15.

- Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Cheung, and Siva Reddy. 2022. [Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 926–937, Dublin, Ireland. Association for Computational Linguistics.
- Xing Han Lù. 2024. [Bm25s: Orders of magnitude faster lexical search via eager sparse scoring](#). *Preprint*, arXiv:2407.03618.
- Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. [Effective transfer learning for identifying similar questions: Matching user questions to covid-19 faqs](#). *Preprint*, arXiv:2008.13546.
- Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. 2020. [Linear mode connectivity in multitask and continual learning](#). *Preprint*, arXiv:2010.04495.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. [Mteb: Massive text embedding benchmark](#). *arXiv preprint arXiv:2210.07316*.
- Clara Na, Ian Magnusson, Ananya Harsh Jha, Tom Sherborne, Emma Strubell, Jesse Dodge, and Pradeep Dasigi. 2024. [Scalable data ablation approximations for language models through modular training and merging](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21125–21141, Miami, Florida, USA. Association for Computational Linguistics.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. [Large dual encoders are generalizable retrievers](#). *Preprint*, arXiv:2112.07899.
- Qwen Team. 2026. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2021. [Searching for scientific evidence in a pandemic: An overview of trec-covid](#). *Preprint*, arXiv:2104.09632.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. [Blended rag: Improving rag \(retriever-augmented generation\) accuracy with semantic search and hybrid query-based retrievers](#). In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, page 155–161. IEEE.
- Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen tau Yih, Pang Wei Koh, and Luke Zettlemoyer. 2025. [Reasonor: Training retrievers for reasoning tasks](#). *Preprint*, arXiv:2504.20595.
- Chaitanya Shivade. 2017. [Mednli — a natural language inference dataset for the clinical domain](#).
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Flax Sentence Embeddings Team. 2021. [Stack exchange question pairs](#).
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- 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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *CoRR*, abs/1807.03748.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. [Embeddinggemma: Powerful and lightweight text representations](#). *Preprint*, arXiv:2509.20354.
- Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. [Retrieval of the best counterargument without prior topic knowledge](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational*

- Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024b. [Searching for best practices in retrieval-augmented generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736, Miami, Florida, USA. Association for Computational Linguistics.
- Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2024. [Promptriever: Instruction-trained retrievers can be prompted like language models](#). *Preprint*, arXiv:2409.11136.
- Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. [Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time](#). *ArXiv*, abs/2203.05482.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Xing. 2024. [LM-cocktail: Resilient tuning of language models via model merging](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2474–2488, Bangkok, Thailand. Association for Computational Linguistics.
- Xing Han Lu. 2024. [publichealth-qa \(revision 3b67b6b\)](#).
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. [Approximate nearest neighbor negative contrastive learning for dense text retrieval](#). In *Proceedings of the 9th International Conference on Learning Representations*.
- Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. [BMRetriever: Tuning large language models as better biomedical text retrievers](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22234–22254, Miami, Florida, USA. Association for Computational Linguistics.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: resolving interference when merging models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. [Language models are super mario: Absorbing abilities from homologous models as a free lunch](#). *Preprint*, arXiv:2311.03099.
- Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. [Optimizing dense retrieval model training with hard negatives](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512. ACM.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.

A Implementation Details

Hyperparameter	
Maximum Tokens	512
Optimizer	AdamW
LR Scheduler	Linear Warmup
# Warmup Steps	100
bf16	True
Learning Rate	
Mistral Instruct v0.2 (7B)	1×10^{-5}
Phi4 mini instruct (3.8B)	2×10^{-5}
Gemma (2B)	1×10^{-5}
Qwen3 (0.6B)	5×10^{-5}
# Epochs	1
LoRA Rank	16
LoRA α	32
Total Batch Size	
Mistral Instruct v0.2 (7B)	32
Phi4 mini instruct (3.8B)	64
Gemma (2B)	128
Qwen3 (0.6B)	256

Table 9: Fine-tuning Hyperparameters.

Model	Size	Avg Med	Avg General	Avg All
BM25	-	0.650	0.774	0.702
Contriever	150M	0.664	0.798	0.720
E5 Large V2	335M	0.724	0.830	0.769
GTR T5 XL	1.2B	0.690	0.814	0.742
BMRetriever	2B	0.691	0.768	0.723
LLM2Vec	3B	0.730	0.846	0.778
Promptriever	7B	0.729	0.831	0.771
<i>STM_{Qwen3}</i>	0.6B	0.713	0.811	0.754
<i>STM_{Gemma}</i>	2B	0.709	0.786	0.741
<i>STM_{Phi4}</i>	3.8B	0.728	0.844	0.776
<i>STM_{Mistral}</i>	7B	0.741	0.856	0.790

Table 10: Retrieval performance measured in Recall@100 across medical and general domains. **Bold** indicates the best result per column.

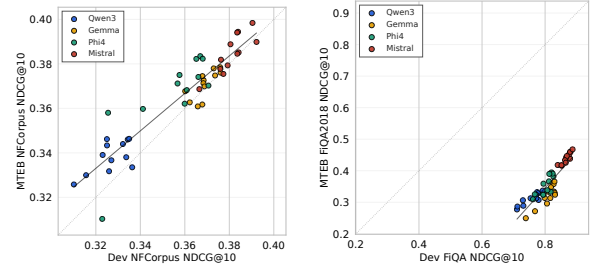
B Prompt Examples

B.1 Hard Negative Generation Prompts

The following prompt template was used for hard negative mining:

Hard Negative Generation Prompt

Given this query: "{query}"
Original prompt: "{original prompt}"
Original positive document: "{positive doc}"
Original negative document: "{original negative}"
Generate a HARD negative document that is:
1. Related to the same domain (extract domain from the query)
2. Contains similar terminology and concepts as the positive document
3. But is NOT relevant to answering the specific query
4. Should be moderately challenging to distinguish from the positive document
5. Should be a realistic document in the domain
6. Should be harder than the original negative document but not as hard as a super hard negative
IMPORTANT: Return ONLY the document text. Do not include any introductory text, explanations, summaries, or meta-commentary. Just return the raw document content.



(a) NFCorpus (medical). (b) FiQA2018 (general). Pearson $r = +0.93$, Spearman $\rho = +0.91$. Pearson $r = +0.89$, Spearman $\rho = +0.89$.

Figure 4: Same-task dev sampling validation (125-query dev NDCG@10 vs. full test benchmark). Each point is one random search model (Qwen3, Gemma, Phi4, Mistral; twelve merge-method variants per architecture). Dotted: identity; solid: linear fit.

C Detailed Results

C.1 Frozen development benchmark

To avoid overfitting merge hyperparameters to the twelve-task MTEB benchmark, we construct a frozen dev split: 500 queries (125 per subset, seed 42) from four retrieval tasks spanning biomedical in-domain retrieval (NFCorpus, CLUE-IR) and general out-of-domain settings (FiQA2018, FEVER). Hyperparameters are selected by un-

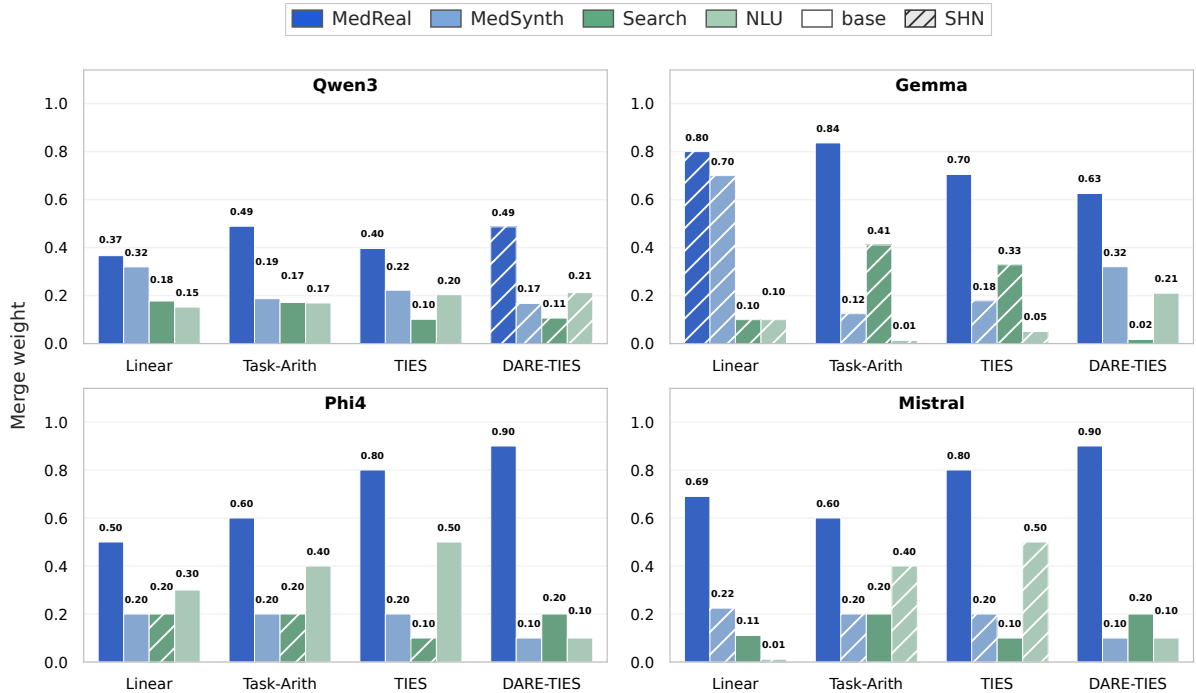


Figure 5: **Expert merge weights** for the best merged model per merging method on each backbone (Qwen3, Gemma, Phi4, Mistral). Each panel shows parent coefficients assigned to four domain experts: MedReal, MedSynth, Search, and NLU. For every backbone and method, we select the merged checkpoint with the highest mean NDCG@10 over the benchmark suite. Solid bars use the standard expert; hatched bars (//) use the synthetic hard-negative (SHN) variant. Bar labels report merge weights.

weighted mean NDCG@10 across the four subsets (Table 14). For NFCorpus and FiQA2018, dev queries are sampled from the official validation split and are disjoint from the full test corpora used in MTEB evaluation; CLUE-IR and FEVER are held exclusively for tuning and do not appear in the benchmark.

Dev-set validity. Dev scores correlate strongly with full MTEB rankings on both overlapping tasks across $n=48$ random-search checkpoints (four architectures $\times$ twelve merge-method variants per architecture; Evolve models excluded, as they optimize directly on dev): Pearson $r=+0.93$ [$+0.89$, $+0.96$] on NFCorpus and $r=+0.89$ [$+0.82$, $+0.94$] on FiQA2018. See Figure 4. Table 23 reports the gap between the best sampled configuration and the arithmetic mean across the $N=10$ random-search draws for each backbone model. Linear merging exhibits a small gap (at most 0.015 across all backbones), indicating that most sampled configurations lie close to the best. Interference-aware methods show wider gaps, up to 0.125 for Phi4 DARE-TIES, reflecting a broader range of sampled configurations.

C.2 Evaluation Benchmark Description

Table 13 describes the twelve retrieval tasks used for evaluation, drawn from the MTEB benchmark (Muennighoff et al., 2022). All tasks are English passage retrieval tasks evaluated with NDCG@10, spanning biomedical and general domains with diverse query types and corpus sizes.

C.3 Merge parameters

We tune merge weights on a shared frozen development benchmark (Table 14) using two complementary search strategies: *random search* (uniform sampling + best-of- N) and *evolve* (CMA-ES). Both use the same twelve variant buckets per architecture: four merge methods (Linear, Task Arithmetic, TIES, DARE-TIES) $\times$ three expert pools: base (four standard fine-tuned experts), hn (four hard-negative experts), and best (a per-architecture mix of standard and HN adapters derived from our strongest evolve runs).

Random search. For each bucket we draw $N=10$ merge configurations with seed 42: weights (and TIES/DARE-TIES densities) are sampled from $\{0.1, 0.2, \dots, 1.0\}$, this yields $12 \times 10 = 120$

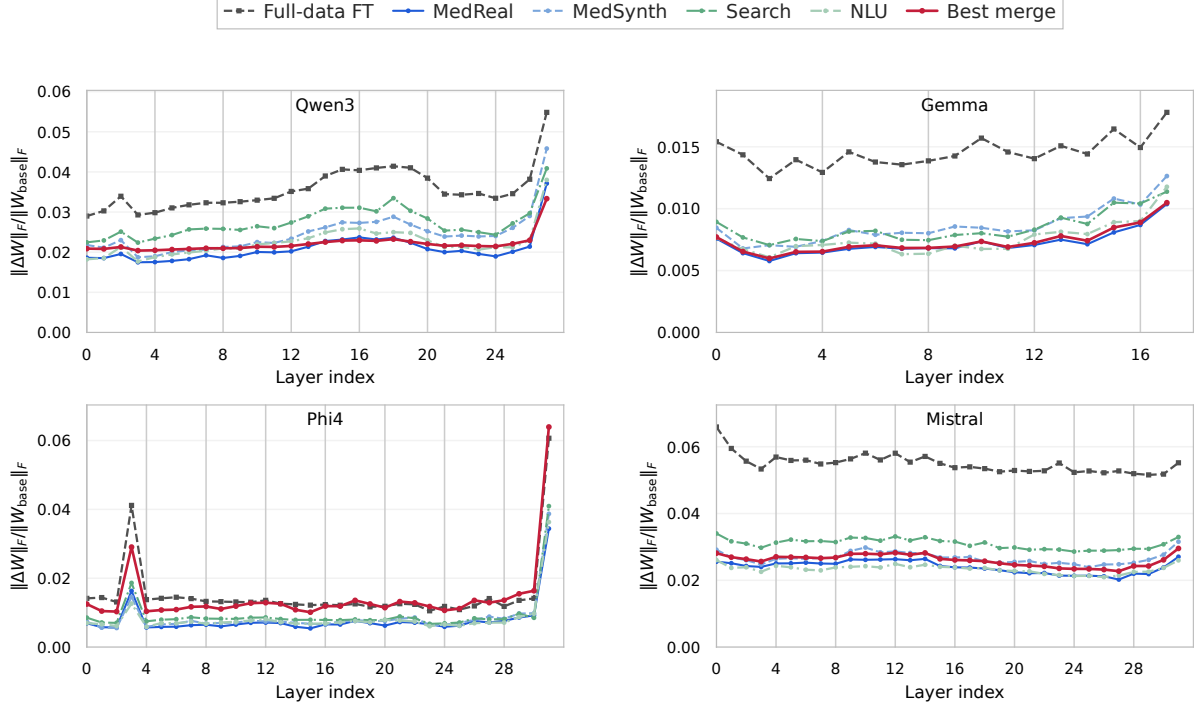


Figure 6: Per-layer relative Frobenius weight change $\|\Delta W\|_F / \|W_{\text{base}}\|_F$ for all four backbone families (Qwen3, Gemma, Phi4, Mistral), where W_{base} is the pretrained base checkpoint. Dashed gray: full-data fine-tuning. Solid red: best merged model (highest mean NDCG@10 across 12 MTEB tasks per family). Colored curves: individual domain experts used in that merged checkpoint.

candidates per architecture; we keep the checkpoint with highest mean dev NDCG@10 per bucket (12 final models per architecture).

Evolve (CMA-ES). We run `mergekit-evolve` with up to 50 function evaluations per bucket. Each evaluation merges four experts, scores the result on the frozen dev split, and returns the mean of four dev NDCG@10 scores to CMA-ES. For Qwen3, best uses the same four HN adapters as hn, but we still run separate buckets and sample independent weights in random search.

Table 16 lists, for each model family, the single best merged checkpoint under *evolve* (CMA-ES on the dev set) and under *random search* (best of ten sampled configurations per expert bucket), each ranked by the mean NDCG@10 in 4 dev tasks.

Method	Qwen3			Gemma			Phi4			Mistral		
	Avg Med	Avg Gen	Avg All	Avg Med	Avg Gen	Avg All	Avg Med	Avg Gen	Avg All	Avg Med	Avg Gen	Avg All
Full Dataset Fine-Tuning												
FT-all _{HNM}	0.622	0.548	0.591	0.638	0.548	0.600	0.655	0.573	0.621	0.692	0.603	0.655
FT-all _{SHN}	0.632	0.555	0.600	0.637	0.561	0.605	0.661	0.585	0.629	0.676	0.609	0.648
Modular Specialized Experts												
Med-Real	0.609	0.510	0.568	0.625	0.542	0.591	0.643	0.567	0.611	0.678	0.616	0.652
Med-Real _{SHN}	0.613	0.541	0.583	0.609	0.527	0.575	0.626	0.564	0.600	0.676	0.606	0.647
Med-Synth	0.614	0.540	0.583	0.606	0.546	0.581	0.642	0.567	0.611	0.671	0.592	0.638
Med-Synth _{SHN}	0.614	0.561	0.592	0.624	0.562	0.598	0.640	0.555	0.605	0.670	0.608	0.644
Search	0.527	0.477	0.506	0.523	0.475	0.503	0.515	0.472	0.497	0.638	0.571	0.610
Search _{SHN}	0.577	0.530	0.558	0.510	0.507	0.508	0.617	0.555	0.591	0.598	0.533	0.571
NLU	0.573	0.509	0.546	0.546	0.484	0.520	0.628	0.525	0.585	0.623	0.575	0.603
NLU _{SHN}	0.578	0.543	0.563	0.564	0.552	0.559	0.585	0.539	0.566	0.637	0.596	0.620
Merged Experts (STM)												
STM _{DARE}	0.631	0.577	0.608	0.649	0.567	0.615	0.659	0.604	0.636	0.684	0.626	0.660
STM _{TIES}	0.634	0.571	0.608	0.649	0.572	0.617	0.659	0.590	0.630	0.693	0.639	0.671
STM _{Linear}	0.636	0.574	0.610	0.631	0.576	0.608	0.648	0.589	0.623	0.687	0.627	0.662
STM _{TA}	0.632	0.572	0.607	0.648	0.573	0.617	0.661	0.595	0.634	0.693	0.633	0.668

Table 11: Average NDCG@10 across 12 MTEB tasks. Specialized experts are trained on partitioned subsets of the training data, while FT-all denotes fine-tuning on the full dataset.

Model	Size	Medical							General					Avg Med.	Avg Gen.	Avg All
		Feedback QA	Medical QA	CUREv1	Public HealthQA	NF Corpus	TREC COVID	Sci Fact	SCI DOCS	Nano FEVER	Argu Ana	Nano Quora	FiQA 2018			
BM25	-	0.563	0.458	0.355	0.718	0.321	0.623	0.686	0.158	0.809	0.492	0.863	0.251	0.532	0.515	0.525
Contriever	150M	0.505	0.592	0.351	0.694	0.313	0.448	0.655	0.171	0.794	0.484	0.944	0.274	0.508	0.533	0.519
E5 Large V2	335M	0.704	0.699	0.562	0.856	0.372	0.666	0.722	0.205	0.889	0.464	0.912	0.411	0.654	0.576	0.622
GTR T5 XL	1.2B	0.577	0.692	0.507	0.713	0.333	0.601	0.642	0.157	0.846	0.528	0.957	0.442	0.581	0.586	0.583
BMRetriever	2B	0.587	0.727	0.471	0.812	0.347	0.839	0.729	0.186	0.940	0.356	0.960	0.357	0.645	0.560	0.609
LLM2Vec	3B	0.708	0.731	0.490	0.814	0.385	0.572	0.746	0.190	0.864	0.553	0.953	0.423	0.635	0.597	0.619
Promptriever	7B	0.714	0.734	0.574	0.855	0.368	0.769	0.759	0.185	0.896	0.526	0.967	0.451	0.682	0.605	0.650
STM _{Qwen3}	0.6B	0.662	0.702	0.473	0.821	0.346	0.762	0.688	0.191	0.824	0.567	0.959	0.327	0.636	0.574	0.610
STM _{Gemma}	2B	0.629	0.728	0.439	0.878	0.379	0.759	0.728	0.205	0.840	0.534	0.954	0.335	0.648	0.573	0.617
STM _{Phi4}	3.8B	0.695	0.747	0.475	0.858	0.382	0.721	0.732	0.225	0.880	0.558	0.964	0.394	0.659	0.604	0.636
STM _{Mistral}	7B	0.737	0.782	0.571	0.895	0.385	0.732	0.751	0.222	0.924	0.629	0.951	0.468	0.693	0.639	0.671

Table 12: Comparison of retrieval performance on English tasks from MTEB. Results are reported primarily on the medical subset, with additional evaluation on five general-domain subsets. Performance is measured using NDCG@10.

Dataset	Domain	#Q	#Docs
TREC-COVID	Biomedical	50	171K
NFCorpus	Biomedical	323	3.6K
SciFact	Biomedical	300	5K
CUREv1	Clinical	2K	245K
PubHealthQA	Biomedical	172	172
MedicalQA	Biomedical	2048	2048
FeedbackQA	Biomedical	1.9K	2.4K
FiQA-2018	Finance	648	57K
ArguAna	General	1.4K	8.7K
SciDocs	Science	1K	25K
NanoFEVER	General	50	5K
NanoQuora	General	50	5K

Table 13: Twelve MTEB evaluation tasks. All tasks use NDCG@10. #Q = number of queries, #Docs = corpus size (approximate).

Subset	Domain	Queries	Docs	Split
NFCorpus	Biomedical	125	2,588	dev
CLUE-IR	Biomedical	125	140	dev
FiQA	General	125	380	dev
FEVER	General	125	202	dev
Total		500	3,310	

Table 14: Frozen dev split used for hyperparameter selection (125 queries per subset, seed 42; corpora contain all relevant documents plus 20% hard negatives).

Method	Pool	Qwen3	Gemma	Phi4	Mistral
DARE-TIES	base	0.603	0.615	0.636	0.660
	shn	0.608	0.604	0.612	0.655
	best	0.606	0.615	0.616	0.659
Linear	base	0.610	0.605	0.621	0.657
	shn	0.607	0.608	0.610	0.643
	best	0.606	0.603	0.623	0.662
Task-Arith	base	0.607	0.614	0.625	0.667
	shn	0.605	0.589	0.616	0.653
	best	0.604	0.617	0.634	0.668
TIES	base	0.608	0.615	0.609	0.658
	shn	0.603	0.611	0.623	0.656
	best	0.606	0.617	0.630	0.671

Table 15: Mean NDCG@10 by expert pool and merge method (max over CMA-ES evolve and random search). Bold: best pool per column within each method block. base: standard experts; shn: synthetic hard-negative experts; best: architecture-specific mix. For Qwen3[†], shn and best use the same four SHN adapters but are searched independently.

Family	Method	Var.	Evolve (CMA-ES)					Random Search				
			Med-Real	Med-Synth	Search	NLU	NDCG @10	Med-Real	Med-Synth	Search	NLU	NDCG @10
Qwen3 [†]	Linear	base	0.37	0.32	0.18	0.15	0.610	hn	1.00	0.40	0.50	0.607
	Task-Arith	base	0.49	0.19	0.17	0.17	0.607	best	0.60	0.20	0.40	0.604
	TIES	base	0.40	0.22	0.10	0.20	0.608	best	0.80	0.20	0.10	0.596
	DARE-TIES	hn	0.49	0.17	0.11	0.21	0.608	base	0.90	0.10	0.20	0.599
Gemma	Linear	base	0.56	0.09	0.20	0.14	0.601	hn	0.80	0.70	0.10	0.608
	Task-Arith	best	0.84	0.12	0.41	0.01	0.617	best	0.70	0.10	0.30	0.616
	TIES	best	0.70	0.18	0.33	0.05	0.617	hn	0.90	0.20	0.30	0.611
	DARE-TIES	base	0.63	0.32	0.02	0.21	0.615	best	0.60	0.50	0.20	0.615
Phi4	Linear	best	0.22	0.31	0.17	0.31	0.623	best	0.50	0.20	0.20	0.623
	Task-Arith	base	0.43	0.25	0.06	0.24	0.625	best	0.60	0.20	0.20	0.634
	TIES	hn	0.52	0.27	0.20	0.05	0.623	best	0.80	0.20	0.10	0.630
	DARE-TIES	base	0.37	0.10	0.24	0.22	0.624	base	0.90	0.10	0.20	0.636
Mistral	Linear	best	0.69	0.22	0.11	0.01	0.662	best	1.00	0.30	0.40	0.655
	Task-Arith	hn	0.41	0.44	0.01	0.19	0.653	best	0.60	0.20	0.20	0.668
	TIES	best	0.77	0.19	0.13	0.06	0.665	best	0.80	0.20	0.10	0.671
	DARE-TIES	best	0.53	0.19	0.16	0.10	0.659	base	0.90	0.10	0.20	0.660

Table 16: Best Evolve (CMA-ES) and Random Search merged checkpoint per architecture and merge method. NDCG@10 is the mean across 12 MTEB tasks. **Var.:** expert pool, base (standard adapters), hn (hard-negative adapters), best (family-specific mix). [†]For Qwen3, hn and best use the same four SHN adapters but are searched independently.

Model	Params (B)	Global BS	# Pairs (M)	FLOPs (approx.)	Δ vs PT+FT
Qwen3 (PT+FT)	0.6	256	11.4	1.26×10^{20}	—
Qwen3 (FT 1.4M)	0.6	256	1.4	1.55×10^{19}	−88%
Gemma (PT+FT)	2.0	128	11.4	4.20×10^{20}	—
Gemma (FT 1.4M)	2.0	128	1.4	5.16×10^{19}	−88%
Phi4 (PT+FT)	3.8	64	11.4	7.98×10^{20}	—
Phi4 (FT 1.4M)	3.8	64	1.4	9.81×10^{19}	−88%
Mistral (PT+FT)	7.0	32	11.4	1.47×10^{21}	—
Mistral (FT 1.4M)	7.0	32	1.4	1.81×10^{20}	−88%

Table 17: Training compute comparison across models and setups. FLOPs are estimated as $3 \times 6ND$ (Kaplan et al., 2020), where N is the number of model parameters, D is the total number of tokens processed (number of pairs $\times$ tokens per pair), and the factor of 3 accounts for the three forward passes required by the contrastive loss. Δ denotes the relative reduction in compute of FT compared to PT+FT.

Variant	Method	Qwen3		Gemma		Phi4		Mistral	
		N@10	R@100	N@10	R@100	N@10	R@100	N@10	R@100
Expert MedReal	–	0.568	0.742	0.591	0.731	0.611	0.766	0.652	0.784
Expert MedSynth	–	0.583	0.738	0.581	0.720	0.611	0.758	0.638	0.777
Expert Search	–	0.506	0.698	0.503	0.665	0.497	0.699	0.610	0.762
Expert NLU	–	0.546	0.719	0.520	0.686	0.585	0.749	0.603	0.763
Expert MedReal (HN)	–	0.575	0.744	0.583	0.727	0.600	0.763	0.647	0.788
Expert MedSynth (HN)	–	0.592	0.742	0.598	0.731	0.605	0.763	0.644	0.780
Expert Search (HN)	–	0.558	0.728	0.508	0.668	0.591	0.753	0.571	0.729
Expert NLU (HN)	–	0.563	0.730	0.559	0.710	0.566	0.750	0.620	0.759
Merged (base experts)	Linear	0.610	0.754	0.605	0.741	0.621	0.773	0.657	0.785
Merged (base experts)	Task-Arith	0.607	0.755	0.614	0.738	0.625	0.764	0.667	0.789
Merged (base experts)	TIES	0.608	0.755	0.615	0.743	0.609	0.765	0.658	0.786
Merged (base experts)	DARE-TIES	0.603	0.753	0.615	0.741	0.636	0.776	0.660	0.787
Merged (HN experts)	Linear	0.607	0.752	0.608	0.739	0.610	0.768	0.643	0.785
Merged (HN experts)	Task-Arith	0.605	0.750	0.589	0.730	0.616	0.767	0.653	0.788
Merged (HN experts)	TIES	0.603	0.749	0.611	0.735	0.623	0.773	0.657	0.786
Merged (HN experts)	DARE-TIES	0.608	0.753	0.604	0.731	0.612	0.769	0.655	0.787

Table 18: Performance comparison across backbone models. HN denotes experts trained with GPT-5-generated hard negatives. N@10 and R@100 denote Avg NDCG@10 and Avg Recall@100, respectively.

Variant	Qwen3	Gemma	Phi4	Mistral
Sum of 4 experts (base)	16.5	22.2	37.3	75.6
Sum of 4 experts (HN)	15.1	19.9	31.9	54.9
Full-data FT (base)	18.7	24.4	83.6	68.8
Full-data FT (HN)	33.4	23.3	36.4	40.5

Table 19: Wall-clock training time (hours) on a single NVIDIA H100.

Backbone	Params compared	Conflict rate
Qwen3 0.6B	596,049,920	87.06%
Gemma 2B	1,981,808,640	86.78%
Phi4 3.8B	3,221,225,472	86.59%
Mistral 7B	6,979,321,856	87.43%

Table 20: Sign-conflict rate across expert task vectors $\tau_k = \theta_k - \theta_B$, computed as the fraction of parameter positions where at least one expert disagrees in sign. Conflict is substantial and consistent across all four families, motivating the use of interference-aware merging methods (TIES, DARE-TIES) that explicitly resolve sign disagreements via majority voting.

Judge	NO (%)	Uncertain (%)	YES (%)
DeepSeek-V4-Flash	96.21	0.40	3.39
Qwen3.6-35B-A3B	95.01	0.20	4.79

Table 21: Independent LLM audit of GPT-5 synthetic hard negatives on a stratified random sample of ($n = 500$) triples. Two judges not used in STM training, DeepSeek-V4-Flash (DeepSeek-AI et al., 2026) and Qwen3.6-35B (Qwen Team, 2026), independently label each candidate as YES (false negative; a correct answer to the query), NO (valid hard negative), or Uncertain. Raw inter-judge agreement is 96.0%, with Cohen’s $\kappa = 0.48$.

Candidate type	Mean	Median	Q25	Q75
sim(positive)	0.729	0.741	0.674	0.798
sim(GPT-5 negative)	0.561	0.573	0.479	0.651
sim(mined negative)	0.539	0.548	0.475	0.613
sim(random negative)	0.202	0.197	0.123	0.270

Table 22: Cosine similarity to the query across candidate types on a stratified sample of ($n=10,000$) triples, measured with Qwen3-Embedding-0.6B (Zhang et al., 2025), an embedding model not used in STM training. GPT-5 hard negatives lie slightly above mined negatives, well below positives, and far above random negatives; the expected regime for a genuine hard negative, topically related to the query but not a correct answer.

Backbone	Linear	Task-Arith	TIES	DARE-TIES
Qwen3	0.005	0.064	0.083	0.087
Gemma	0.006	0.036	0.038	0.029
Phi4	0.008	0.095	0.107	0.125
Mistral	0.015	0.048	0.059	0.062

Table 23: Sweep range of dev-set NDCG@10 across the $N=10$ random-search configurations per (backbone, method) bucket, aggregated across the three expert pools ($n=30$ per cell). Values report the gap between the best sampled configuration (Max) and the arithmetic mean of the ten. Linear merging is largely insensitive to the sampled weights (≤ 0.015 across backbones); interference-aware methods (TIES, DARE-TIES) show wider sweep ranges (up to 0.125), reflecting a stronger dependence on the specific weight configuration.