Multi-Stage Contrastive Learning with Joint Domain-Specific Masked Supervision for Domain Adaptation of Sentence Embedding Models

Anonymous ACL submission

Abstract

We present a multi-stage contrastive learning framework for domain adaptation of sentence embedding models, incorporating joint domain-specific masked supervision. Our approach addresses the challenges of adapting large-scale general-domain sentence embedding models to specialized domains. By jointly optimizing masked language modeling (MLM) and contrastive objectives within a unified training pipeline, our method enables effective learning of domain-relevant representations while preserving the robust semantic discrimination properties of the original model. We empirically validate our approach on both high-resource and low-resource domains, achieving improvements up to 13.4% in NDCG@10 over strong general-domain baselines. Comprehensive ablation studies further demonstrate the effectiveness of each component, highlighting the importance of balanced joint supervision and staged adaptation.

1 Introduction

Self-supervised learning has enabled significant progress in natural language processing, with methods like MLM (Devlin et al., 2019; Liu et al., 2020; Conneau et al., 2020; Sanh et al., 2019) and contrastive training (Reimers and Gurevych, 2019; Wu et al., 2020; Liu et al., 2021; Yan et al., 2021) driving recent developments. However, these methods are typically explored separately, as effectively combining MLM and contrastive learning remains a significant challenge, since their joint optimization often results in conflicting training signals and suboptimal performance (Gao et al., 2021). Nevertheless, unifying these objectives presents an opportunity to leverage the complementary strengths of token-level (MLM) and sentence-level (contrastive) supervision, while also improving the quality of learned representations by mitigating the anisotropy problem (a phenomenon that confines embeddings to a narrow cone-like region in the vector space, thereby limiting their expressiveness) (Ethayarajh, 2019; Li et al., 2020; Gao et al., 2021). While there have been successful attempts to combine MLM and contrastive objectives for training language models (Meng et al., 2021; Chi et al., 2021) and sentence embeddings (Gao et al., 2021; Wu et al., 2022; Giorgi et al., 2021), the majority of the prior work has focused on general-domain data.

General-domain sentence embedding models are now widely available, many trained on vast general-domain corpora using a two-stage approach: an initial pre-training phase on massive unlabeled data, followed by supervised fine-tuning (Wang et al., 2022b; Li et al., 2023; Nussbaum et al., 2024; Merrick et al., 2024). The data used for pre-training can exceed half a billion sentence pairs (hundreds of gigabytes of text), resources that are rarely available in specific domains. Although these general-domain models can perform competitively in specialized areas, their lack of domain-specific knowledge often limits performance. To address this gap, we propose domain adaptation of pre-trained embedding models that leverage their ability to distinguish between similar and dissimilar pairs and transfer it to a domain-specific embedding model.

Previous research on language model adaptation highlights the importance of domain-specific vocabulary for improving results on downstream tasks (Beltagy et al., 2019; Gu et al., 2020). However, simply adding domain-specific vocabulary and continuing MLM training degrades the contrastive properties of the learned representations, since the encoder loses its desirable characteristics under the token prediction objective (Wu et al., 2022). On the other hand, adding new tokens and continuing only with the contrastive objective provides insufficient training signals to update new domain tokens, as the embedding matrix receives diluted signals due to the pooling functions applied to generate sentence embeddings.

This dilemma motivates our approach of using a joint objective to enable both token-level and sentence-level supervision, thus benefiting from both worlds and enhancing domain adaptation for both the encoder and the embedding matrix during training. Building on a mutual information maximization perspective (Hjelm et al., 2018; Bachman et al., 2019; Kong et al., 2019; Chen et al., 2020; Chi et al., 2021), which demonstrates that these objectives are aligned rather than contradictory, operating at different levels of language granularity, we leverage the joint optimization of MLM and contrastive objectives. Though these objectives are theoretically aligned, a key challenge in joint training arises from the dominance of the MLM loss and more frequent token-level supervision, which can overwhelm the joint objective and hinder balanced optimization. This issue can be mitigated by carefully controlling the strength of the MLM signal during joint training, directing it to domain-relevant signals, without resorting to encoder separation, which may limit the propagation of informative token-level signals into sentence-level representations.

To thoroughly evaluate our method, we apply it to both high-resource and low-resource domains. Most domain-adaptation research focuses on high-resource and medium-resource domains, which is valuable for benchmarking, comparison with strong baselines, and conducting ablation studies. Yet this focus restricts the generalizability of adaptation methods to truly low-resource domains, which are common in real-world applications. Such domains often face acute data scarcity, making robust adaptation methods essential for ensuring equitable access to state-of-the-art language technologies and maximizing the real-world impact of embedding models. To demonstrate the robustness and practical value of our approach, we validate it in two domains: the Biomedical domain, which is characterized by high-resource scientific texts, and the Islamic domain, which represents low-resource but culturally significant content. This allows us to test the robustness of our method even when there is very limited in-domain data.

Our main contributions are as follows: (1) We propose a novel domain adaptation approach for pretrained sentence embedding models that jointly optimizes MLM and contrastive objectives within a mutual information maximization framework. (2) We empirically validate our method on both high-resource (biomedical) and low-resource (Islamic) domains, demonstrating substantial gains over strong general-domain baselines, with up to a 2.8% average improvement in NDCG@10 across biomedical benchmarks and a 13.3% improvement in NDCG@10 on the Islamic dataset. (3) We conduct comprehensive ablation studies to analyze the contribution of each component and the dynamics of joint objective training. (4) We release our code and pretrained models to support reproducibility and facilitate future research.

2 Related Work

2.1 Contrastive and MLM Objectives

Contrastive Predictive Coding (CPC) is one of the earliest works to introduce the InfoNCE loss (van den Oord et al., 2018). The loss encourages informative representations to align with a given anchor while distinguishing them from negative examples. Though this work is not specific to sentence embeddings, it laid the theoretical groundwork for contrastive learning. **SimCSE** targeted sentence embeddings by proposing a contrastive learning setup utilizing dropout noise to generate two distinct views of the same sentence, thereby optimizing a contrastive loss between them (Gao et al., 2021). In their appendix, the authors mentioned an experiment involving the incorporation of MLM during training. However, they found that performance dropped, likely due to a conflict between MLM's token-level loss and the sentence-level contrastive signal. Built on this challenge in multi-objective training when using a shared encoder, the authors of **InfoCSE** introduced a more sophisticated framework (Wu et al., 2022). Rather than combining MLM and contrastive loss on the same encoder output, InfoCSE uses an auxiliary lightweight encoder. This architectural separation prevents MLM gradients from interfering with the contrastive training of final-layer CLS embeddings. InfoCSE showed improved performance over SimCSE in STS benchmarks. Conversely, the authors of **DeCLUTR** explicitly combine MLM with contrastive training for sentence embeddings (Giorgi et al., 2021). They construct positive pairs from contiguous spans of the same document and apply a standard BERT-style masking on the anchor span and train jointly through a single encoder. Evaluation on SentEval benchmarks for classification and similarity showed that a unified objective is a promising approach. **COCO-LM** integrates

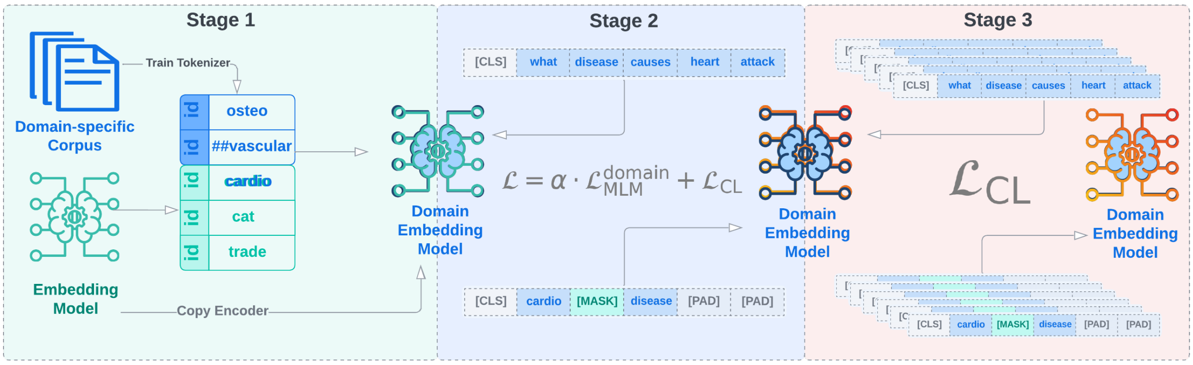
Stage 1
Stage 2
Stage 3
Domain-specific Corpus
Train Tokenizer
id osteo
id ##vascular
id cardio
id cat
id trade
Embedding Model
Copy Encoder
Domain Embedding Model
[CLS] what disease causes heart attack
$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{MLM}}^{\text{domain}} + \mathcal{L}_{\text{CL}}$
[CLS] cardio [MASK] disease [PAD] [PAD]
Domain Embedding Model
$\mathcal{L}_{\text{CL}}$
Domain Embedding Model

Figure 1: Multi-stage domain adaptation of sentence embedding models.

contrastive learning into a pretraining pipeline for transformer language models (Vaswani et al., 2017; Meng et al., 2021) and replaces BERT's Next Sentence Prediction (NSP) objective with a more effective contrastive signal, pairing corrupted and truncated versions of a sentence. Corruption is performed using an ELECTRA-style generator (Clark et al., 2020), producing fluent but subtly altered sequences. The model jointly learns to align these pairs (via contrastive loss) and to correct the corruption (through token-level denoising). COCO-LM demonstrated consistent gains on GLUE tasks (Wang et al., 2018), showing that contrastive objectives outperform NSP for general-purpose pretraining. The authors of **InfoXLM** reframe masked language modeling as a contrastive prediction task, formulating it with the InfoNCE loss (Chi et al., 2021). When combined with a sentence-level cross-lingual contrastive objective, this joint training enables InfoXLM to achieve state-of-the-art results on cross-lingual understanding and retrieval benchmarks.

2.2 Domain Adaptation

Domain adaptation is most commonly performed at the language modeling stage, where general-purpose models undergo continued pre-training on in-domain corpora (Lee et al., 2019; Alsentzer et al., 2019). Such approaches typically suffer from the absence of domain-specific vocabulary, which often necessitates training from scratch (Beltagy et al., 2019; Gu et al., 2020). To avoid these GPU-heavy methods, recent work has explored lightweight domain adaptation by introducing new domain vocabulary to already well-trained models, thereby expediting the pre-training process (Poerner et al., 2020; Sachidananda et al., 2021; Pavlova and Makhlouf, 2023). In contrast to our approach, these strategies have primarily been applied to language models prior to downstream task training. For sentence embedding models that have already undergone contrastive training, domain adaptation efforts have mostly focused on data-driven approaches such as data augmentation, denoising objectives, or generative pseudo-labeling (Thakur et al., 2021; Wang et al., 2021, 2022a). In our work, we focus on a model-driven approach.

3 Multi-stage Contrastive Learning with Domain-Specific Masked Supervision

3.1 Augmenting Contrastive Models with Domain-Specific Vocabulary

To leverage the robust encoder learned during contrastive pretraining, we reuse both the encoder and the original embedding matrix. However, to accommodate a word distribution shift from a general domain vocabulary to a new domain vocabulary, we augment the model with new domain-specific tokens (see Figure 1):

Domain-Specific Tokenizer Training. We begin by training a new tokenizer on a large domain-specific corpus to identify vocabulary units that capture relevant terminology.

Domain Vocabulary Augmentation. We then identify domain-specific tokens that are missing from the original tokenizer used by the contrastive model, and incorporate these into the model's embedding matrix, initializing their embeddings as the average of their base model subword embeddings.

This design choice is motivated by the fact that contrastive training mainly shapes the encoder. By modifying only the input vocabulary, we retain the original encoder weights from the pretrained contrastive model, preserving its sentence-level discrimination capabilities.

3.2 Joint Optimization of Contrastive and MLM Objectives

Jointly optimizing MLM and contrastive objectives can theoretically combine the benefits of fine-grained token-level supervision from MLM with sentence-level supervision encouraged by contrastive learning. However, in practice, it is difficult to perform joint optimization on both. Below, we detail the reasoning behind this challenge and propose our approach to balance these objectives effectively. To motivate our approach, we start with the information-theoretic interpretation of both MLM and contrastive objectives (Chi et al., 2021). Both objectives can be viewed as maximizing a mutual information lower bound. Using the InfoNCE formulation from van den Oord et al. (2018), the contrastive objective for context pairs c_1 and c_2 can be expressed as:

$$I(c_1; c_2) \geq \mathop{\mathrm{E}}_{q(\mathcal{N})}\left[\log \frac{f_\theta(c_1, c_2)}{\sum_{c' \in \mathcal{N}} f_\theta(c_1, c')}\right] + \log|\mathcal{N}| \tag{1}$$

where f_θ is a scoring function that measures similarity between two contexts c_1 and c_2 (e.g., via dot product or cosine similarity) and $\mathcal{N}$ represents a set of negative contexts.

Similarly, MLM can also be interpreted as maximizing a mutual information lower bound between the context c_1 and the masked token x_1 with $\mathcal{N}$ being the vocabulary:

$$I(c_1; x_1) \geq \mathop{\mathrm{E}}_{q(\mathcal{N})}\left[\log \frac{f_\theta(c_1, x_1)}{\sum_{x' \in \mathcal{N}} f_\theta(c_1, x')}\right] + \log|\mathcal{N}| \tag{2}$$

The InfoNCE formulation highlights that while two objectives may be aligned, there is a significant imbalance between them. The larger vocabulary size in MLM results in a substantially larger denominator, leading to very low probabilities for the correct token. Consequently, this generates higher loss values and, therefore, larger gradient magnitudes, causing MLM to dominate the training process. Moreover, MLM operates at the token level, often producing many more learning signals per batch. This imbalance leads to stronger and more frequent gradients for MLM. As a result, the model disproportionately optimizes the MLM objective, leaving the contrastive component under-trained. To counteract this, we propose restricting MLM to only a subset of the vocabulary — the domain vocabulary, which includes only newly introduced domain-specific tokens. This reduces the size of the denominator by replacing the full vocabulary V_{all} with a smaller domain-specific set V_{domain} limiting the masking signal to rare, informative tokens. Rewriting the MLM loss with domain vocabulary gives:

$$\begin{aligned} & I(c_1; x_1) \\ & \geq \mathop{\mathrm{E}}_{q(\mathcal{N_D})}\left[\log \frac{f_\theta(c_1, x_1)}{\sum_{x' \in \mathcal{N_D}} f_\theta(c_1, x')}\right] + \log|\mathcal{N_D}| \end{aligned} \tag{3}$$

In this variant, the set $\mathcal{N_D}$ contains only domain-specific vocabulary tokens. This targeted vocabulary reduction refocuses the MLM objective on domain-critical tokens, providing clearer and less overpowering gradient signals, which align more closely with those of the contrastive objective.

While the InfoNCE form provides theoretical grounding, in practice both MLM and contrastive learning are usually implemented using cross-entropy losses. For the contrastive loss, this takes the form:

$$\mathcal{L}_{\text{CL}} = -\log \frac{\exp\left(\phi(c_1)^\top \phi(c_2)\right)}{\sum_{c' \in \mathcal{N}} \exp\left(\phi(c_1)^\top \phi(c')\right)} \tag{4}$$

where $\phi(\cdot)$ is an encoder that maps the input to a dense vector, and $\mathcal{N}$ includes one positive and $|\mathcal{N}| - 1$ negatives.

Similarly, the domain-focused MLM cross-entropy loss becomes:

$$\mathcal{L}_{\text{MLM}}^{\text{domain}} = -\log \frac{\exp\left(\phi(c_1)^\top e(x_1)\right)}{\sum_{x' \in \mathcal{V}} \exp\left(\phi(c_1)^\top e(x')\right)} \tag{5}$$

Here, ϕ is the shared encoder (same as used in the contrastive loss), e is the embedding lookup table, and $\mathcal{V}$ is the (domain-constrained) candidate token vocabulary.

Our final joint loss is expressed as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{MLM}}^{\text{domain}} + \mathcal{L}_{\text{CL}} \tag{6}$$

where α is a scalar coefficient used to balance the gradient magnitude. This formulation ensures that both objectives contribute to optimizing the shared encoder while mitigating the gradient dominance of MLM. Ultimately, by limiting the MLM's vocabulary set and calibrating its contribution to the joint objective, our approach effectively integrates the strengths of MLM and contrastive training, resulting in robust and domain-adaptive representations. It is crucial to highlight that in our design, the inputs to contrastive objectives are provided with mask perturbation, which forces the model to disambiguate which specific tokens distinguish negative documents from positive (see Figure 1). In this way, MLM acts as a localized supervision signal that highlights the differences and similarities between pairs, particularly in cases where contrastive loss alone may struggle due to mean pooling or similar functions, which average over token embeddings and blur these distinctions. By applying MLM-guided masking, the model learns to focus on the key differentiating features.

3.3 Contrastive-Only Training

For the third stage, we continue training our model using only the contrastive objective, after the new domain tokens have been introduced and learned. This stage serves as a corrective step, allowing the encoder to recover and reinforce sentence-level discrimination, which may be diluted during joint MLM+contrastive training. By focusing solely on contrastive learning, the model re-aligns its representations to produce robust sentence embeddings.

4 Experiments on a High-Resource Domain

Training Data. To construct a large-scale biomedical corpus, we parsed the 2025 PubMed snapshot and extracted *(title, abstract)* pairs. When available, metadata such as journal name and keywords were appended to the title to enrich the context. We filtered out non-English entries as well as pairs where either the title or abstract was too short to form a meaningful sentence pair. To further ensure data quality and minimize false positives, we applied a consistency-based filtering procedure using the `gte-base` model (see Appendix A). This resulted in approximately 20 million high-quality sentence pairs for use in stages two and three of our approach. We evaluate our models in a zero-shot setting. To avoid any risk of benchmark data leakage (a common issue with sentence embedding models), we fine-tune on BioASQ Task 9a (Tsatsaronis et al., 2015). This dataset consists essentially of human-selected PubMed title–abstract pairs (approximately 16 million), each annotated with MeSH (Medical Subject Headings) that we append to titles to form our queries. This data is used for fine-tuning the final model after the third stage of our training pipeline.

Evaluation Data and Metrics. We evaluate on the medical subset of the MTEB (Massive Text Embedding Benchmark) (Muennighoff et al., 2022), a standardized benchmark for assessing the quality of text embeddings across a diverse set of tasks, such as retrieval, classification, clustering, reranking, semantic textual similarity (STS), and summarization. We use BiorxivClusteringP2P, MedrxivClusteringP2P, and MedrxivClusteringS2S for clustering (V-measure); MedicalQARetrieval, NFCorpus, SciFact, and TRECCOVID for retrieval (nDCG@10); and BIOSSES for STS (Spearman correlation). We report results on BIOSSES in Table 1, but the analysis on the STS task is performed as a part of the ablation Section 4.2.

Baselines. For unsupervised baselines, we use nomic-embed-text-v1$_{\text{unsup}}$[1] as our primary baseline representing an unsupervised contrastive embedding model pretrained on general-domain data. We also train this model on the unsupervised training data described above and include the `nomic-embed-bio` model in the comparison. To analyze the impact of each stage on domain adaptation, we use three models from our pipeline: **Biomedical Initial**, which adds new domain-specific vocabulary to the contrastive model without further pretraining (Stage 1); **Biomedical-Joint MLM+Contrastive (BJMC)**, trained with both masked language modeling and contrastive objectives on domain data (Stage 2); and **Biomedical Contrastive Only (BCO)**, further trained with the contrastive objective alone (Stage 3). For supervised baselines, we select a diverse set of well-established embedding models that report MTEB scores on biomedical datasets, as listed on the official MTEB leaderboard[2].

Implementation Details. We implement the joint MLM and contrastive training on top of the Nomic repository[3]. For the purely contrastive stage, we reuse the original implementation from the repository. The model architecture is based on BERT (Devlin et al., 2019) with several modifications introduced by the Nomic repo. At the first

[1] https://huggingface.co/nomic-ai/nomic-embed-text-v1-unsupervised

[2] https://huggingface.co/spaces/mteb/leaderboard

[3] https://github.com/nomic-ai/contrastors/tree/main

Model	**BIOSSES**	**BiorxivC**	**MedicalQAR**	**MedrxivP2P**	**MedrxivS2S**	**NFCorpus**	**SciFact**	**TRECCOVID**
				Unsupervised models				
nomic-embed-text-v1$_{\text{unsup}}$	87.189	38.78	68.307	34.854	**32.521**	**35.684**	71.982	62.203
nomic-embed-bio	87.012	36.107	66.173	30.72	28.552	34.235	73.302	**62.203**
Biomedical Initial	78.946	33.747	63.58	30.261	25.126	26.091	67.246	57.050
BJMC	**88.116**	38.101	68.677	34.536	29.882	32.217	72.535	60.763
BCO	88.057	**39.31**	**70.233**	**35.089**	30.287	34.137	**74.710**	61.281
				Supervised models				
E5$_{\text{base}}$ (Wang et al., 2022b)	85.103	37.49	68.051	34.6347	32.0616	36.589	73.083	79.638
GTE$_{\text{base}}$ (Li et al., 2023)	87.642	40.62	71.455	36.404	**34.9025**	**37.897**	**76.178**	68.783
BGE$_{\text{base}}$ (Xiao et al., 2023)	85.533	-	-	-	-	35.539	73.258	76.447
text-embedding-ada-002	86.351	-	-	-	-	36.972	72.746	68.474
nomic-embed-text-v1	86.471	41.48	66.648	37.0082	34.3009	35.028	70.500	**79.923**
Bio-embed-model	**89.869**	**42.551**	**72.378**	**37.865**	32.631	35.571	75.875	63.546

Table 1: Evaluation of unsupervised and supervised models across biomedical benchmarks. Bold indicates the highest score per column within each group.

stage, we add approximately 9k new biomedical tokens. We set the masking rate to 0.15, the MLM loss weighting hyperparameter $\alpha = 0.3$ throughout the joint training phase. Details of ablation on α and masking rate can be found in Section 4.2, and other hyperparameter settings are provided in Appendix C. We train stages two and three of the proposed pipeline using only in-batch negatives, and additionally include hard-mined negatives during fine-tuning.

4.1 Results and Analysis

Our results demonstrate several important trends regarding domain adaptation for sentence embeddings (see Table 1). First, we observe that simply continuing pretraining a general domain embedding model (`nomic-embed-text-v1`$_{\text{unsup}}$ (Nussbaum et al., 2024)) on in-domain data can lead to reduced performance compared to the original general-domain baseline across most benchmarks (as in the `nomic-embed-bio` model), suggesting that naive in-domain adaptation may distort learned representations. This issue becomes even more pronounced when augmenting the vocabulary with domain-specific tokens without any retraining (**Biomedical Initial**), resulting in a substantial performance drop across all datasets, likely due to embedding mismatch. In contrast, our multi-stage approach consistently restores and enhances performance: applying a joint MLM+contrastive objective (**BJMC**) recovers and further improves results, while a final contrastive-only training stage (**BCO**) achieves the highest scores on four benchmarks (BiorxivClusteringP2P, MedicalQARetrieval, MedrxivClusteringP2P, and SciFact), resulting in the best average performance overall and with a 2.8% increase over the general-domain baseline. These results highlight a clear progression across adaptation stages, where naive vocabulary expansion leads to degradation, targeted joint supervision restores model quality, and a final contrastive stage enables robust domain adaptation. In the supervised setting, our **Bio-embed-model** achieves the highest scores on BiorxivClusteringP2P, MedrxivClusteringP2P, and MedicalQARetrieval, outperforming other strong supervised baselines on these key biomedical tasks. We also observe that the largest improvements are seen in clustering and STS tasks, indicating that domain-adapted masked supervision is particularly beneficial for capturing fine-grained semantic relationships and latent structure in biomedical texts. Overall, the improvements are robust across tasks and settings, demonstrating the practical value of our approach for real-world biomedical and specialized text retrieval scenarios. However, as shown in Table 1, all of our models lag on the TRECCOVID dataset (Voorhees et al., 2021). Inspection of the TRECCOVID queries reveals that, alongside core biomedical and clinical questions, a significant fraction focuses on social or policy aspects of the pandemic (approximately 20%). Such queries, addressing societal impacts or interventions like school reopening, may fall outside the primary scope of biomedical corpora used for model adaptation. This mismatch in domain coverage could partly explain the observed underperformance. Moreover, recent large-scale analyses of PubMed using embedding-based atlases have shown that COVID-19 literature forms a uniquely isolated cluster in embedding space, with strong internal topical fragmentation, further challenging biomedical models (Kobak et al., 2024).

Model	Score
BJMC	**88.116**
BJMC (All-Token MLM)	63.995
BJMC (mlm_prob 0.3)	49.871
BJMC (alpha 0.1)	76.032
BJMC (alpha 0.2)	81.336
BJMC (alpha 0.4)	86.794
BJMC (alpha 0.5)	67.708
BJMC (as the 3d stage)	70.540
Contrastive only	84.428

Table 2: Performance comparison of BJMC and ablated models on BIOSSES.

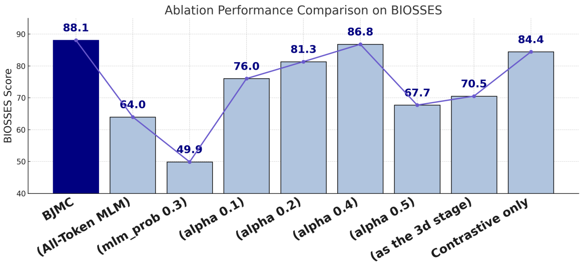

Figure 2: BIOSSES score for various ablation settings in unified model study.

4.2 Ablation Studies

The primary focus of our ablation study is the second stage of the proposed method (joint MLM+contrastive training). Accordingly, all ablation experiments are conducted on the **Biomedical-Joint MLM+Contrastive** (**BJMC**) model. For ablation, we use BIOSSES (Soganciogl u et al., 2017), an STS dataset that requires models to capture fine-grained semantic relationships between sentences, beyond what is assessed in standard retrieval or clustering tasks (Cer et al., 2017); this enables us to demonstrate the effect of the MLM objective.

Effect of Masking Strategy. To evaluate the effectiveness of domain-restricted masked language modeling (MLM), we compared our default approach, which restricts MLM to domain-specific tokens, with an alternative that applies MLM to all vocabulary tokens (All-Tokens MLM). This change led to a 27% decrease in performance (see Table 2), highlighting the critical importance of directing the masking signal towards domain-specific terms.

Masking rate. We further ablate the effect of the masking rate by increasing the MLM probability from the default 0.15 to 0.3 during joint training. As shown in Table 2, raising the masking rate leads to a dramatic drop in performance (from 88.1 to 49.9), indicating that excessive masking can overwhelm the contrastive signal and degrade the learned representations.

Alpha hyperparameter. We also ablate the effect of the MLM loss weight (α), which controls the relative contribution of the MLM objective during joint training. We systematically explore a range of α values—a hyperparameter whose impact is rarely examined in prior literature, despite its crucial role in balancing objectives. As shown in Table 2 and Figure 2, setting $\alpha = 0.5$ causes the MLM loss to dominate, resulting in a drastic performance drop. At the other extreme, $\alpha = 0.1$ does not sufficiently promote learning of new domain tokens, and $\alpha = 0.2$ yields only modest gains. While $\alpha = 0.4$ remains competitive though slightly suboptimal, the highest performance is achieved at $\alpha = 0.3$, indicating it as the most balanced choice for our joint objective.

Order of Training Stages. Next, we reverse the order of stages 2 and 3 by first performing only contrastive training with a large batch, followed by contrastive training combined with MLM. As shown in Table 2, this results in a noticeable performance drop from 88.116 to 70.548, a decrease of 20%. This suggests that applying the joint objective to an already strong embedding model can disturb its contrastive capability.

No Joint Objective. Finally, we assess the impact of the joint MLM+contrastive objective by removing the second stage entirely and training solely with the contrastive objective after vocabulary expansion (“Contrastive only”). As shown in Table 2, omitting the MLM stage results in a performance drop from 88.1 to 84.4, indicating that joint training with domain-restricted MLM provides a meaningful boost over contrastive adaptation alone.

As shown in Figure 2, increasing the MLM probability to 0.3, increasing the α weight to 0.5, or applying MLM masking to all tokens causes the most severe performance drops, demonstrating that excessive MLM signal overwhelms joint training and degrades representation quality.

5 Experiment on a Low-Resource Domain

Experimental Setup. As noted in Section 1, the Islamic domain is a low-resource area, especially for English-language data. In-domain data suitable for training sentence embedding models is extremely scarce. To address this, we constructed an in-domain training set by extracting semantically related verse pairs from *Tafseer Ibn Kathir* and

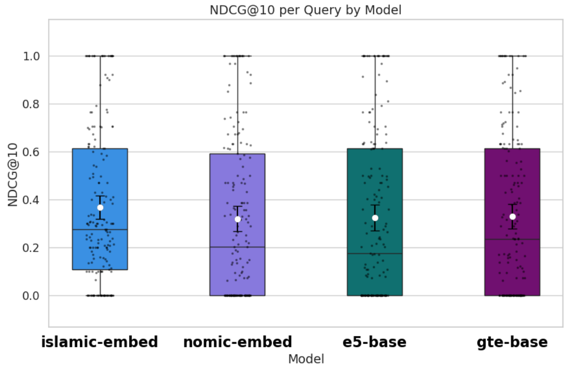

Figure 3: Boxplot of per-query NDCG@10 scores for all models.

Model	NDCG@10
Islamic-embed-model	**36.809**
GTE$_{\text{base}}$	32.924*
E5$_{\text{base}}$	32.466*
nomic-embed-text-v1	32.048**

Table 3: NDCG@10 evaluation results on the Islamic dataset. * indicates statistical significance at $p < 0.1$ and ** at $p < 0.05$ (paired t-test vs. Islamic-embed-model).

applying consistency filtering with the `gte-base` model, resulting in 7,587 high-quality pairs. Further details on the data construction process are provided in Appendix B.

Although the Islamic domain in English is characterized by an extremely limited amount of available training data, it is notable for having a dedicated evaluation dataset—unlike many other low-resource domains. Recent efforts by Malhas and Elsayed (2020) have created a verified high-quality Qur'anic Reading Comprehension Dataset (QRCD), which includes questions frequently asked within the Islamic domain. The answers provided are exhaustive, meaning all Qur'anic verses directly responding to the questions have been thoroughly extracted and annotated. To increase the size of the evaluation set, we combine the training and development splits, resulting in a total of 169 queries for testing. Although QRCD is originally in Arabic, we employ verified English translations to enable evaluation in the English language. For retrieval collections, we use the Sahih International English translation.[4] We compare our final model with three strong general-domain embedding models using NDCG@10 as the evaluation metric. The implementation details follow those used for the biomedical model, with the following modifications: we add 3k domain-specific tokens to the vocabulary.

Results. The **Islamic-embed-model** achieves the highest NDCG@10 score (36.8; Table 3). All models exhibit considerable variation in per-query scores (Figure 3), reflecting the challenging nature of the dataset, but our model's upper quartile and mean are both higher. Notably, the lower whisker for the **Islamic-embed-model** does not reach the minimum value of 0, whereas the lower whiskers for the general-domain models extend to 0. This indicates that our model makes fewer completely incorrect predictions (i.e., queries with NDCG@10 = 0), while the comparison models sometimes fail to retrieve any relevant results for certain queries. The upper whiskers are similar across all models, suggesting comparable best-case performance, but the reduction in low and zero scores for our model contributes to its higher overall mean NDCG@10. This performance gap can be attributed to differences in pretraining data: while biomedical content constitutes a measurable minority of large-scale pretraining corpora (Wang et al., 2022b; Li et al., 2023; Nussbaum et al., 2024), Islamic domain texts are almost absent (typically less than 0.01%). This negligible coverage leaves general-domain models ill-equipped to capture the linguistic and conceptual nuances of Islamic texts, making domain adaptation essential for low-resource areas.

6 Conclusion

We present a novel approach for domain adaptation of sentence embedding models by jointly optimizing MLM and contrastive objectives. Unlike standard domain adaptation methods, which are typically applied at the language modeling stage or after task-specific training via data augmentation, our method leverages a model-driven approach for domain adaptation after contrastive training. We achieve robust gains in both high-resource (biomedical) and low-resource (Islamic) domains, surpassing general-domain baselines even with limited in-domain data.

Limitations

Much of the research on domain adaptation focuses on high-resource domains such as biomedicine, where data is abundant and benchmarks are well established. In this work, we explicitly include a low-resource domain (Islamic text), recognizing both

[4] `https://tanzil.net/trans/`

the additional challenges and the importance of extending language technologies to underrepresented settings. However, we recognize that each domain, whether high- or low-resource, can present unique characteristics and challenges that could affect the effectiveness of domain adaptation methods. As such, the generalizability of our approach may vary depending on domain-specific linguistic features, data availability, or cultural context. We encourage further research on adaptation strategies that are sensitive to the specific requirements and risks of diverse domains.

Ethical Considerations

Adapting models to specialized domains may amplify biases or inaccuracies present in domain-specific corpora. For example, biomedical texts may reflect publication biases or outdated medical practices, while religious texts may encode culturally specific viewpoints. In our work, we rely exclusively on publicly available and verified resources for data collection and model training; no private or proprietary data is used at any stage. Nevertheless, we acknowledge that these sources may still carry implicit biases or limitations. We encourage users of domain-adapted models to consider these factors carefully, especially when applying the models in sensitive or high-impact contexts. The models will be released under the Apache-2.0 license to ensure transparency, reproducibility, and broad accessibility. The model `nomic-embed-text-v1`$_{\text{unsup}}$ is licensed under Apache-2.0. All artifacts used in this study are open-source and available for research purposes. We utilized AI assistants for debugging, optimizing LaTeX formatting, and improving grammar clarity.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Neural Information Processing Systems*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

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.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *ArXiv*, abs/1808.06670.

Dmitry Kobak, Dong He, et al. 2024. The landscape of biomedical research. *bioRxiv*.

Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. 2019. A mutual information maximization perspective of language representation learning. *ArXiv*, abs/1910.08350.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yining Li, Yuhui Zhang, Xiaoman Pan, Yinan Li, Ning Ding, Wei Wu, Yujing Wang, Xiaoyan Zhu, and Minlie Huang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1–10, Online. Association for Computational Linguistics.

Rana Malhas and Tamer Elsayed. 2020. Ayatec: building a reusable verse-based test collection for arabic question answering on the holy qur'an. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(6):1–21.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul N. Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. In *Neural Information Processing Systems*.

Logan Merrick, Cameron Beeler, Mike Lewis, Rohit Girdhar, David Hall, Xilun Chen, John Thickstun, and Jacob Devlin. 2024. Arctic-embed: Scalable, efficient, and accurate text embedding models. *arXiv preprint arXiv:2405.05374*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Vera Pavlova and Mohammed Makhlouf. 2023. BIOptimus: Pre-training an optimal biomedical language model with curriculum learning for named entity recognition. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 337–349, Toronto, Canada. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and covid-19 QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Vin Sachidananda, Jason Kessler, and Yi-An Lai. 2021. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language*

Processing, pages 155–165, Virtual. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Gizem Soganciogl u, Arzucan Ozgur, and Hakime Ozturk. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Jens Petersen, Jörg Hakenberg, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. In *BMC bioinformatics*, volume 16, page 138. BioMed Central.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William Hersh, Kirk Roberts, and Lucy Lu Wang. 2021. Trec-covid: Constructing a pandemic information retrieval test collection. *Journal of the American Medical Informatics Association*, 28(4):766–773.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022a. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. InfoCSE: Information-aggregated contrastive learning of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3060–3070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian yun Nie. 2023. C-pack: Packed resources for general chinese embeddings. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

A Consistency-based Filtering Procedure

To further ensure data quality and minimize false positive pairs, we employed a semantic filtering procedure using the `gte-base` model. Specifically, we first sampled up to 1 million candidate query–document pairs from the initial dataset. Each query and document was independently encoded into dense vector representations using the `gte-base` sentence embedding model.

Next, we constructed a FAISS index from all document embeddings to enable efficient similarity search. For each query embedding, we retrieved the top-k most similar document embeddings from the index, based on cosine similarity. If the original paired document d_i was not found among the top-k retrieved documents for its corresponding query q_i, we discarded the pair (q_i, d_i). This filtering step ensures that only pairs with strong semantic alignment—according to the embedding model—are retained for further training.

The intuition behind this approach is to eliminate weakly related or noisy pairs that may have been erroneously grouped together in the initial data extraction. By keeping only those pairs where the document is highly ranked for its query, we improve the quality and relevance of training examples, leading to better domain adaptation during model training.

B Curating Passages for Training the Islamic Domain Model

Dense retrieval models often experience performance degradation when applied to new domains, emphasizing the value of training on in-domain data. The scarcity of such data is typically addressed through augmentation techniques like synthetic data generation, paraphrasing, pair recombination, round-trip translation, or denoising autoencoders. However, these approaches risk altering the original semantics, which is especially problematic for sensitive religious and heritage texts. To overcome this, we utilize Tafseer Ibn Kathir, a classical and authoritative Qur'anic exegesis rich in verse commentary and inter-verse references. This resource enables natural and semantically meaningful augmentation of training data by explicitly linking related verses.

Pair Extraction. Let C_t denote the collection of Tafseer texts by Ibn Kathir. We extract all verse pairs $V_t = (v_q, v_p)$ referenced in C_t, resulting in approximately 11,000 candidate pairs.

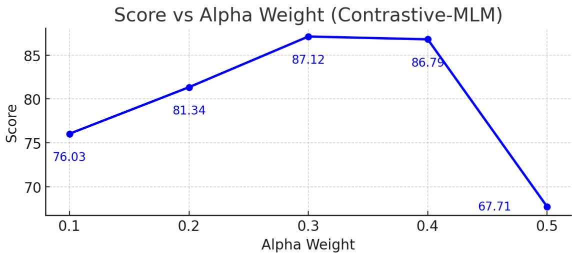

Figure 4: Effect of alpha weight on the performance in the 2nd stage of Contrastive+MLM training.

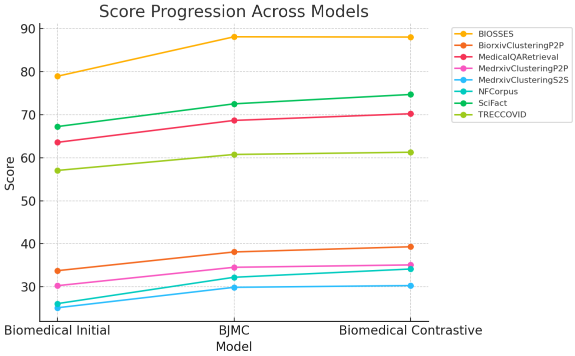

Figure 5: The impact of each stage on the MTEB datasets.

Filtering. Not all extracted pairs represent strong semantic correlations suitable for retrieval training, due to indirect or implicit relationships. To select high-quality positive pairs, we score each candidate (v_q, v_p) using the `gte-base` model to obtain similarity scores $s = \texttt{gte-base}(v_q, v_p)$. Pairs scoring below a predefined threshold are removed, yielding a filtered set V_f of 7,587 robust positive pairs for training.

C Training Hyperparameters

Computing Infrastructure	1x H100 (80 GB)
Hyperparameter	**Assignment**
number of epochs	1-5
batch size	128-49k
sequence length	64-256
maximum learning rate	0.0005
learning rate optimizer	Adam
learning rate scheduler	None or Warmup linear
Weight decay	0.01
Warmup proportion	0.06
learning rate decay	linear

Table 4: Hyperparameters for training and finetuning sentence embedding models.