LASER: LLM-GUIDED SEVERITY-AWARE REPORT ENCODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unstructured radiology reports contain rich clinical information, yet converting them into structured representations that capture the severity of findings remains challenging. Pre-trained language models compress specialized clinical text into a small region of their feature space, causing clinical severity distinctions, such as discrete versus massive calcification, to collapse into indistinguishable representations. We propose LASER, a label-efficient framework that reshapes the latent space of a text encoder to explicitly separate clinical concepts and impose a continuous severity ordering, without any manual annotation. An LLM generates synthetic sentences for each clinical concept–severity combination, providing structured training signal at zero labeling cost. A multi-objective training scheme separates clinical concepts and builds a continuous severity ordering within each concept, amplifying distinctions compressed by generic pre-training. We evaluate LASER on two datasets across two languages and organ systems, a German cardiac CT cohort and the English CheXpert Plus chest X-ray dataset Chambon et al. (2024), on severity classification and report retrieval. LASER consistently outperforms the baselines on both datasets and tasks, demonstrating that LLM-guided synthetic supervision is an effective and scalable path toward structured, interpretable clinical text representations.

1 INTRODUCTION

Radiology reports are a cornerstone of clinical documentation, capturing detailed observations about pathological findings and their severity in free-text form. Automatically extracting structured information from these reports, for example the severity of clinical findings, holds significant potential for downstream applications such as risk stratification, disease monitoring, and clinical decision support. A natural approach is to leverage pre-trained language models to encode report sentences into latent representations. However, pre-trained text encoders are trained on a large corpus of texts, where specialized reports only get a very small area in the feature space, leading to difficulties in distinguishing clinically relevant aspects like the severity of a diagnostic finding. Encoding this gradient of severity is critical for two reasons: first, for discriminative performance, as collapsing severity distinctions removes the signal that separates a patient with massive findings from one with discrete findings; second, for interpretability, as without explicit concept separation and severity ordering, the position of an embedding in the feature space carries no clinically meaningful information.

To address this, we propose LASER, a label-efficient framework that trains a concept- and severity-aware language encoder on LLM-generated synthetic sentences without any manual annotation. The synthetic sentences are used as geometric anchors that explicitly define the target structure of a clinical embedding space, encoding both concept identity and severity ordinality without any real labeled data. Unlike prior work that uses LLM-generated data for classification or general-purpose embedding training, LASER shapes the latent geometry itself, enabling severity estimation and report retrieval on real-world radiology reports via nearest-neighbor lookup in the structured space. Our main contributions are:

- A label-efficient pipeline using LLM-generated synthetic sentences to train a concept- and severity-aware encoder.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

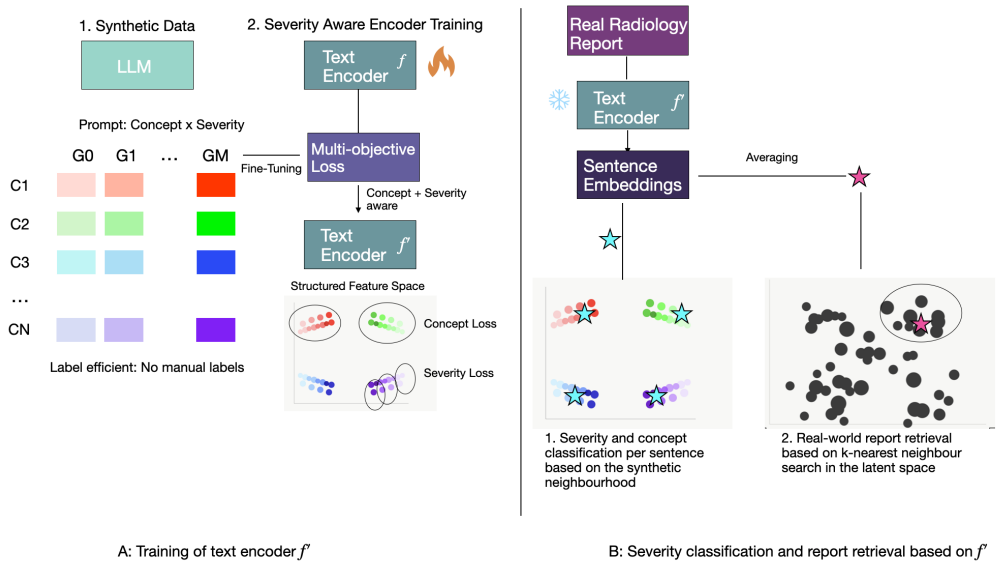


Figure 1: Schematic illustration of LASER. A: Training of the text encoder f'. 1. An LLM generates synthetic sentences for each combination of concept and severity without manual annotation. 2. A text encoder is fine-tuned to learn a structured concept- and severity-aware feature space. B: Severity/Concept classification and report retrieval. The trained model f' generates sentence embeddings from a real radiology report. 1. The severity and concept per sentence are predicted by looking at the k-nearest neighborhood of labeled synthetic sentences in the latent space. 2. The sentence embeddings of one report are averaged to generate one single embedding used for report retrieval. Similar real-world radiology reports are retrieved by looking at the k-nearest neighborhood in the structured latent space of the averaged sentence embeddings.

- A multi-loss training objective combining concept classification, contrastive learning, hard negative mining, triplet loss, and cosine regression to enforce concept separation and continuous severity ordering
- A concept- and severity-aware text encoder with a structured latent space, usable for severity classification and report retrieval on real-world data.
- Extensive ablation studies analyzing the contribution of each loss term, the influence of using contrastive loss functions as ordinal geometric supervision compared to categorical classification and the performance of our multi-objective loss compared to the state-of-the-art ordinal Rank-N-Contrast loss (RNC) Zha et al. (2023).

We evaluate LASER on two datasets across two languages and organ systems, demonstrating consistent improvements over state-of-the-art text encoders. LASER shows that LLM-generated synthetic sentences can serve not merely as training data, but as codebook vectors that anchor specific concept–severity regions in the latent space. This geometric supervision induces an ordinally structured feature space that generalizes to real radiology reports. An overview of the whole architecture is shown in Fig.1.

2 Related Work

Pre-trained Language Models: BERT Devlin et al. (2019) established large-scale pre-training for natural-language processing (NLP) and has been widely adopted for clinical text. Domain specific models like ClinicalBERT Alsentzer et al. (2019) address the vocabulary shift between general and biomedical language. German BERT Chan et al. (2020) and German MedBERT Bressem et al. (2024) provide domain-adapted pre-training for German clinical text. More recently, large-scale multilingual sentence encoders such as E5 Wang et al. (2024a) have been developed to produce semantically meaningful sentence-level representations via contrastive pre-training on large text cor-

pora, achieving strong performance on semantic similarity tasks across languages. Despite these advances, none of the models explicitly encode the ordinal structure of clinical severity within their feature spaces, as their pre-training objectives do not account for fine-grained clinical gradations.

Synthetic Training Data: Recent work has demonstrated that LLM-generated synthetic data can be used to train high-quality text embedding models without any manually labeled data. Wang et al. (2024b) generates synthetic data with an LLM for a large number of text embedding tasks across different languages to fine-tune open-source decoder-only LLMs. Similarly, Lee et al. (2024) distills knowledge from LLMs into a compact retriever by generating synthetic query-passage pairs and refining them via LLM-based relabeling. While these approaches demonstrate the power of LLM-generated data for general-purpose embedding training, they optimize for semantic similarity across diverse tasks without imposing explicit structural constraints on the embedding space. In contrast, LASER uses LLM-generated synthetic sentences not as general-purpose training data, but as geometric anchors that define a clinically structured target geometry, thereby explicitly encoding both concept identity and ordinal severity in the latent space.

Metric Learning and Structured Latent Spaces: To impose structure on latent representations, metric learning methods such as contrastive learning Chopra et al. (2005) and the triplet loss Schroff et al. (2015) learn embeddings, by pulling similar ones together and separating dissimilar ones. Supervised contrastive learning Khosla et al. (2020) extends this by using explicit label information. Rank-N-Contrast Zha et al. (2023) further proposes to learn continuous representations for regression by contrasting samples based on their rankings in the target space, guaranteeing that the learned representations respect the continuous ordering of labels. Ordinal representation learning has been explored in computer vision for tasks such as age estimation and disease grading, where monotonic constraints or ordinal regression losses are used to impose label ordering on the embedding space Lim et al. (2019); Tang et al. (2023). In the clinical domain, these approaches have been explored for applications such as medical image retrieval and report matching Zhong et al. (2021); Zhang et al. (2025). However, its use for encoding an ordinal severity structure specifically in radiology text remains largely unexplored. LASER bridges this gap by combining supervised contrastive learning, hard negative mining, triplet loss, and cosine regression to jointly enforce concept separation and continuous severity ordering.

3 METHODOLOGY

We train a concept- and severity-aware text encoder that is used for severity classification and report retrieval. The text encoder is pre-trained with synthetic sentences generated by an LLM for each concept and severity, using a multi-objective loss that enforces concept separation and severity ordering. The fine-tuned text encoder is used for severity classification and report retrieval.

3.1 SYNTHETIC SENTENCE GENERATION

We identify clinical concepts that are present in the reports as $\{c_i, ..., c_n\}$, where n is the number of concepts. Additionally, grade/severity levels are defined as $\{g_i, ..., g_m\}$, where m is the number of severities present for the concepts. For each concept + severity, an LLM generates synthetic sentences without any manual labeling, yielding a training set $\mathcal{S}_{\text{train}} = \{(\mathbf{s}_i, c_i, g_i)\}_{i=1}^{P}$ and validation set $\mathcal{S}_{\text{val}} = \{(\mathbf{s}_i, c_i, g_i)\}_{i=1}^{R}$, where each $\mathbf{s}_i$ is associated with a concept label $c_i \in \{1, ..., n\}$ and a grade label $g_i \in \{1, ..., m\}$.

3.2 REPORT ENCODING

We train a report encoder on the synthetic sentences, each labeled with a clinical concept c and severity level g. We fine-tune a pre-trained text encoder f with five loss functions: a Cross-Entropy loss $\mathcal{L}_{Concept}$, a Multiple-Negative-Ranking loss$\mathcal{L}_{MNRL}$, a Hard-Negative loss$\mathcal{L}_{HN}$, a triplet loss$\mathcal{L}_{Tri}$ and a cosine regression lossL_{Cos}. The first one $L_{Concept}$, is a cross-entropy loss for the concept prediction, enforcing concept-separation. The Multiple-Negative-Ranking loss $\mathcal{L}_{MNRL}$ pulls same-concept, same-severity sentences into compact clusters. For the given dataset $\mathcal{V}$, it uses normalized embeddings $\hat{\mathbf{z}}_i$, a temperature parameter τ, pairwise similarities between embeddings $s_{ij} = \hat{\mathbf{z}}_i^\top \hat{\mathbf{z}}_j / \tau$

and positive set $\mathcal{P}(i)$ of same-concept, same-grade sentences:

$$\mathcal{L}_{MNRL} = -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\left(\log\sum_{j\in\mathcal{P}(i)} e^{\tilde{s}_{ij}} \;-\; \log\sum_{k\in\mathcal{V},k\neq i} e^{\tilde{s}_{ik}}\right), \tag{1}$$

where $\tilde{s}_{ij} = s_{ij} - \max_k s_{ik}$ works as a stability shift to prevent numerical overflow during computation. $\mathcal{L}_{HN}$ extends this, by penalizing high similarity between same-concept, different-severity sentences:

$$\mathcal{L}_{\text{HN}} = \frac{1}{|\mathcal{V}|^2}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}}\cos(\mathbf{z}_i,\mathbf{z}_j)\cdot\mathbf{1}\big[c_i = c_j\big]\cdot\mathbf{1}\big[g_i \neq g_j\big], \tag{2}$$

where c_i is the concept label, g_i is the grade label and $\cdot\mathbf{1}$ is the indicator function (hard negative mask). $\mathcal{L}_{Tri}$ Schroff et al. (2015) enforces a relative ordering margin between same- and different-severity embeddings, using same-concept, same-severity sentences as positives and same-concept, different-severity sentences as negatives. $\mathcal{L}_{Cos}$ is a cosine regression loss, enforcing continuous ordering of severity levels within each concept:

$$\mathcal{L}_{\text{Cos}} = \frac{\displaystyle\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}}\mathbf{1}[c_i = c_j]\cdot\Big(\cos(\mathbf{z}_i,\mathbf{z}_j) - t_{ij}\Big)^2}{\displaystyle\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}}\mathbf{1}[c_i = c_j]}, \qquad t_{ij} = 1 - \frac{|g_i - g_j|}{3}, \tag{3}$$

where t_{ij} is the target similarity (1 for the same grade, 0 for the maximum distance) and $\mathbf{1}[c_i = c_j]$ the mask to ensure this loss is only calculated within the same concepts. Together, the four losses enforce compactness, separation, relative ordering, and a continuous ordinality. The five loss functions are weighted and summed $\mathcal{L} = \lambda_{Concept} * \mathcal{L}_{\text{Concept}} + \lambda_{MNRL} * \mathcal{L}_{\text{MNRL}} + \lambda_{HN} * \mathcal{L}_{\text{HN}} + \lambda_{Tri} * \mathcal{L}_{\text{Tri}} + \lambda_{Cos} * \mathcal{L}_{\text{Cos}}$. We get a fine-tuned text encoder f'.

3.3 Severity Prediction

The fine-tuned text encoder f' can be used to classify the severity and concept of a real-world sentence. A new sentence is embedded by f', yielding $\mathbf{v}$. The synthetic validation sentences S_{val} are also embedded, representing codebook vectors for every concept + grade combination. The clinical concept c and severity grade g of the real sentence $mathbfv$ are predicted via k-nearest neighbor search in the structured feature space. The cosine similarity between $\mathbf{v}$ and each codebook vector is computed as:

$$\text{sim}(\mathbf{v},\mathbf{s}_i) = \frac{\mathbf{v}^\top\mathbf{s}_i}{\|\mathbf{v}\|\cdot\|\mathbf{s}_i\|}. \tag{4}$$

The k most similar codebook vectors are retrieved:

$$\mathcal{N}_k(\mathbf{v}) = \underset{i}{\text{top-}k}\ \text{sim}(\mathbf{v},\mathbf{s}_i), \tag{5}$$

and the concept $\hat{c}$ and grade $\hat{g}$ is assigned via majority vote over all k neighbors:

$$\hat{c} = \arg\max_c \sum_{i\in\mathcal{N}_k(\mathbf{v})}\mathbf{1}[c_i = c], \hat{g} = \arg\max_g \sum_{i\in\mathcal{N}_k(\mathbf{v})}\mathbf{1}[g_i = g]. \tag{6}$$

This assignment method directly leverages the structured geometry of the learned feature space.

3.4 Report Retrieval

To evaluate whether the structured feature space captures clinically meaningful similarity between reports, we assess the ability of the text encoder to retrieve reports with similar concept-severity profiles. Given a set of N radiology reports, each report r_j is represented by a mean-pooled embedding $\mathbf{e}_j \in \mathbb{R}^d$, computed by averaging all sentence embeddings of the report. Additionally, we use the concept classification described before, to assign sentences to their respective concept and pooling only embeddings from the same concept together, resulting in a set of embeddings $\{\mathbf{e}_j^{c_i}, ..., \mathbf{e}_j^{c_n}\}$, with one embedding for each concept.

Relevance Determination: For evaluation purposes, each report is characterized by a grade profile $\mathbf{p}_j = (g_j^{c_1}, ..., g_j^{c_n})$, where each entry represents the maximum severity grade observed across all sentences of the respective concept. In case a concept is missing from the report, the grade is set to -1, the pooled concept specific embedding to a zero vector and this concept is not considered during the retrieval. Two reports r_j and r_l are considered relevant to each other if their severity profiles are sufficiently similar. For the overall retrieval evaluation, relevance is defined based on the mean absolute error (MAE) between severity profiles:

$$\text{rel}(r_j, r_l) = \mathbf{1}\left[\frac{1}{|\mathcal{C}_{jl}|}\sum_{c\in\mathcal{C}_{jl}} |g_j^c - g_l^c| \leq \tau\right], \tag{7}$$

where $\mathcal{C}_{jl}$ is the set of concepts that are present in both reports and τ is a similarity threshold, i.e. how much each concept is allowed to deviate across the two reports.
For the concept-specific retrieval, relevance is defined per concept c:

$$\text{rel}_c(r_j, r_l) = \mathbf{1}\left[|g_j^c - g_l^c| \leq \tau\right]. \tag{8}$$

Retrieval: For each report r_j, all other reports are ranked based on the cosine similarity of their embeddings:

$$\text{sim}(\mathbf{e}_j, \mathbf{e}_l) = \frac{\mathbf{e}_j^\top \mathbf{e}_l}{\|\mathbf{e}_j\| \cdot \|\mathbf{e}_l\|}, \tag{9}$$

for the overall retrieval. For the concept-specific retrieval, the concept-specific pooled embeddings are utilized for the cosine similarity ranking:

$$\text{sim}(\mathbf{e}_j^c, \mathbf{e}_l^c) = \frac{\mathbf{e}_j^{c\top} \mathbf{e}_l^c}{\|\mathbf{e}_j^c\| \cdot \|\mathbf{e}_l^c\|}. \tag{10}$$

4 Experiments

4.1 Data

Real Data: For evaluation purposes we utilize two real world radiology report datasets with different languages and organ systems. Firstly, reports from the publicly available CheXpert Plus dataset Chambon et al. (2024) are used, describing findings from 2D chest X-rays. From the documented findings we identify three that are suitable for severity reporting: pulmonary edema, pleural effusion and cardiomegaly. Each of these three concepts has four possible grades: negative, mild, moderate and severe. A subset of 700 reports is selected and the sentences are manually labeled with concept and grade, resulting in 1714 labeled sentences. The second dataset is an internal dataset consisting of 130 reports with 906 sentences from 3D cardiac CT images. We identify four clinical concepts present in our cardiac CT reports: coronary artery calcification, coronary artery stenosis, abnormal myocardium and a dilated aorta or truncus pulmonalis (tp). Each of these four concepts can again have four different grades of severity: negative, mild, moderate and severe. Both datasets were manually labeled concerning concept and grade. In case a sentence does not concern any of the defined concepts or describes more than one of the defined concepts, it is filtered out via keyword search during pre-processing.

Synthetic Data: Synthetic sentences were generated using Claude Anthropic (2026) with 50, 100, 200, and 300 per severity per concept for both datasets. Only concept and severity describing keywords (e.g. stenosis + mild, edema + severe) were provided as prompts, no real reports. Additionally, a validation dataset with 40 sentences per severity per concept was generated to use for the k-nearest neighbor classification on the real sentences. Full prompts to generate the synthetic sentences are given in A.4.2.

4.2 Experimental Details

Report Encoder: The report encoder is initialized with the Multilingual E5-large text encoder Wang et al. (2024a) and a single linear classification head with output dimension 3 for the CheXpert dataset and 4 for the internal cardiac dataset, corresponding to the number of defined concepts.

Training: During training, a custom batch sampler ensures two sentences per severity and concept in every batch, resulting in a batch size of 32 for our internal dataset and 24 for the CheXpert dataset. We use Adam Kingma & Ba (2014), a learning rate of $1e-6$ and 200 epochs. All losses share the same weights ($\lambda_{MNRL} = \lambda_{HN} = \lambda_{Tri} = \lambda_{Cos} = \lambda_{Concept} = 1$) during training. Five training runs with different seeds were conducted. Details on the implementation are given in AppendixA.4.1.

Ablations: We conduct three ablation studies: (i) removing individual loss terms to assess their influence on the performance, (ii) to separate the effect of the proposed approach from supervised training on severity labels, we train the same encoder on the identical synthetic sentences with a cross-entropy severity objective (Classifier) and (iii) evaluating the performance when using the same data for fine-tuning but instead of our multi-objective loss using the state-of-the-art Rank-N-Contrast (RNC) loss Zha et al. (2023). Full results are provided in AppendixA.1.

Baselines: As baselines, we choose state-of-the art text encoders. Since the internal dataset consists of german reports, the german BERT model Chan et al. (2020) and german MedBERT model Bressem et al. (2024) are utilized, whereas for the CheXpert dataset the normal uncased BERT model Devlin et al. (2019) and english BioClinical BERT Alsentzer et al. (2019) are used. For both datasets, we also include the Multilingual E5-large text encoder Wang et al. (2024a) and RadBERT Yan et al. (2022) as baselines.

Evaluation: On both datasets, the concept classification as well as severity classification based on k-nearest neighbor search of synthetic sentences are evaluated. Here, the number for k-nearest neighbors is set to 10. The severity classification is evaluated over all concepts as well as for each concept separately. We evaluate the concept and severity classification in terms of accuracy and F1-score. For the severity classification the quadratic weighted kappa (QWK) is given as well to assess the ordinal ordering. Additionally, a report retrieval task is evaluated. For the overall retrieval we test two thresholds, 0.0 and 0.5, i.e. exact matches and a maximum deviation in terms of severity of 0.5 across the concepts. For the concept specific task, we only test exact matches, i.e. a threshold of 1.0. Here we evaluate the precision, normalized discounted cumulative gain (NDCG) and mean reciprocal rank (MRR). For the internal cardiac dataset the precision and NDCG are given @1, @5 and @10 and for the CheXpert dataset @1, @5, @10 and @20 since we have 700 reports for the CheXpert dataset and 130 reports for the cardiac dataset. To quantify if LASER induces a geometrically structured latent space, we analyze the embedding space of real radiology report sentences. Here, we compute the Spearman-correlation coefficient between the first PCA component of each concept cluster and the corresponding severity grade, to show that the severity is encoded as a continuous geometric axis within the concept cluster.

5 Results

5.1 Concept and Severity Classification

Quantitative results for the concept and severity classification on the internal dataset are shown in Tab.1. It can be seen that all LASER configurations outperform the baselines on the severity classification task in terms of accuracy, F1-score and QWK. This is true for the overall classification as well as for the classifications per concept for every concept. The best performing LASER configuration is LASER_{200}, i.e. 200 sentences per concept during training (50 per severity). Here, an accuracy of 0.8250, F1-score of 0.6971 and QWK of 0.7640 is achieved on the overall category. For the concept classification, the Multilingual E5-large text encoder Wang et al. (2024a) achieves the highest accuracy with 0.9790 and F1-score with 0.9800. Quantitative results for the concept and severity classification on the CheXpert dataset are shown in Tab.2. Again, it can be seen that all LASER configurations outperform the baselines on the severity classification task for each concept as well as overall. The best performing LASER configuration is LASER_{800}. Here, an accuracy of 0.8022, F1-score of 0.6854 and QWK of 0.6324 is achieved on the overall category. Additionally, all LASER configurations outperform the baselines on the concept classification task, where every configuration achieves an accuracy and F1-score higher than 98%. For both datasets, confusion matrices for the severity classification are shown in Appendix A.3.

	Metric	German BERT	German MedBERT	RadBERT	Multilingual E5-large	LASER_{200}	LASER_{400}	LASER_{800}	LASER_{1200}
Concept	Acc	0.9161	0.8720	0.6689	**0.9790**	0.9236 ± 0.0207	0.9435 ± 0.0088	0.9492 ± 0.0056	0.9541 ± 0.0091
	F1	0.9218	0.8915	0.6933	**0.9800**	0.9268 ± 0.0193	0.9443 ± 0.0095	0.9522 ± 0.0055	0.9614 ± 0.0065
Severity: Overall	Acc	0.4007	0.6115	0.6214	0.6689	$\mathbf{0.8250} \pm 0.0164$	0.7854 ± 0.0210	0.7633 ± 0.0134	0.7572 ± 0.0094
	F1	0.3269	0.4622	0.2882	0.5571	$\mathbf{0.6971} \pm 0.0174$	0.6794 ± 0.0234	0.6621 ± 0.0200	0.6553 ± 0.0093
	QWK	0.2652	0.4244	0.0753	0.6653	$\mathbf{0.7640} \pm 0.0636$	0.7390 ± 0.0532	0.7100 ± 0.0488	0.6853 ± 0.0308
Severity: Calcification	Acc	0.4804	0.7696	0.3186	0.7696	0.8304 ± 0.0107	0.8294 ± 0.0088	0.8176 ± 0.0217	$\mathbf{0.8323} \pm 0.0175$
	F1	0.4266	0.6604	0.2101	0.6531	0.7292 ± 0.0156	$\mathbf{0.7215} \pm 0.0107$	0.6884 ± 0.0247	0.6919 ± 0.0275
	QWK	0.3613	0.7010	-0.1682	0.8413	0.9170 ± 0.0048	$\mathbf{0.9134} \pm 0.0034$	0.8885 ± 0.0149	0.8995 ± 0.0090
Severity: Stenosis	Acc	0.2447	0.4918	0.7435	0.5341	$\mathbf{0.8066} \pm 0.0233$	0.7082 ± 0.0361	0.6692 ± 0.0211	0.6348 ± 0.0225
	F1	0.2496	0.3088	0.2897	0.4484	$\mathbf{0.6311} \pm 0.0094$	0.5630 ± 0.0325	0.5456 ± 0.0320	0.5181 ± 0.0121
	QWK	0.2505	0.3297	0.0526	0.5498	$\mathbf{0.6310} \pm 0.1046$	0.5809 ± 0.0871	0.5574 ± 0.0774	0.5128 ± 0.0461
Severity: Myocardium	Acc	0.6240	0.7600	0.6880	0.8320	0.8032 ± 0.0175	0.8448 ± 0.0269	0.8464 ± 0.0215	$\mathbf{0.9104} \pm 0.0255$
	F1	0.3884	0.4125	0.3014	0.5147	0.5982 ± 0.0170	0.6927 ± 0.0422	0.7300 ± 0.0352	$\mathbf{0.8040} \pm 0.0388$
	QWK	0.3712	0.6217	0.3652	0.8045	0.8730 ± 0.0080	0.9097 ± 0.0155	0.8696 ± 0.0338	$\mathbf{0.9341} \pm 0.0085$
Severity: Aorta/TP	Acc	0.5461	0.6118	0.6316	0.7763	0.8868 ± 0.0183	$\mathbf{0.8934} \pm 0.0118$	0.8855 ± 0.0075	0.8717 ± 0.0136
	F1	0.2836	0.3647	0.2780	0.6133	0.7475 ± 0.0418	$\mathbf{0.7731} \pm 0.0288$	0.7557 ± 0.0109	0.6929 ± 0.0712
	QWK	0.1027	0.4411	0.2292	0.6902	0.9179 ± 0.0134	$\mathbf{0.9235} \pm 0.0067$	0.9136 ± 0.0052	0.8966 ± 0.0123

Table 1: Concept and severity classification results on the german cardiac CT dataset. Results are given for the overall concept classification and severity classification over all concepts as well as separate for calcification, stenosis, abnormal myocardium and dilated aorta/tp.

		BERT	BioClinical BERT	RadBERT	Multilingual E5-large	LASER_{200}	LASER_{400}	LASER_{800}	LASER_{1200}
Concept	Acc	0.5974	0.9002	0.9312	0.9632	0.9862 ± 0.0010	0.9859 ± 0.0008	0.9861 ± 0.0018	$\mathbf{0.9879} \pm 0.0017$
	F1	0.5563	0.8954	0.9254	0.9602	$\mathbf{0.9859} \pm 0.0010$	0.9855 ± 0.0008	0.9857 ± 0.0018	0.9875 ± 0.0018
Severity: Overall	Acc	0.3116	0.4889	0.5776	0.4901	0.7079 ± 0.0234	0.7785 ± 0.0134	$\mathbf{0.8022} \pm 0.0121$	0.7909 ± 0.0082
	F1	0.2193	0.3832	0.4599	0.3617	0.5947 ± 0.0369	0.6812 ± 0.0259	$\mathbf{0.6854} \pm 0.0347$	0.6446 ± 0.0208
	QWK	0.0472	0.2391	0.3500	0.2717	0.4586 ± 0.0808	0.6059 ± 0.0318	$\mathbf{0.6324} \pm 0.0376$	0.5816 ± 0.0286
Severity: Edema	Acc	0.4011	0.4492	0.4898	0.4011	0.6946 ± 0.0307	0.7442 ± 0.0028	$\mathbf{0.7472} \pm 0.0128$	0.7316 ± 0.0040
	F1	0.2572	0.2977	0.3045	0.2404	0.5156 ± 0.0576	$\mathbf{0.5599} \pm 0.0078$	0.5280 ± 0.0215	0.4903 ± 0.0058
	QWK	0.0285	0.2346	0.2535	0.1522	0.3924 ± 0.0779	$\mathbf{0.4944} \pm 0.0079$	0.4863 ± 0.0265	0.4263 ± 0.0070
Severity: Effusion	Acc	0.3096	0.5274	0.6281	0.6296	0.7535 ± 0.0248	0.8047 ± 0.0049	$\mathbf{0.8145} \pm 0.0129$	0.8101 ± 0.0059
	F1	0.2270	0.4550	0.6032	0.5858	0.7175 ± 0.0533	$\mathbf{0.7790} \pm 0.0103$	0.7663 ± 0.0318	0.7236 ± 0.0157
	QWK	0.1810	0.2298	0.4118	0.5090	0.5767 ± 0.0663	0.6586 ± 0.0128	$\mathbf{0.6592} \pm 0.0326$	0.6268 ± 0.0139
Severity: Cardiomegaly	Acc	0.2169	0.4799	0.6044	0.3976	0.6607 ± 0.0152	0.7803 ± 0.0389	$\mathbf{0.8422} \pm 0.0254$	0.8293 ± 0.0285
	F1	0.1652	0.3236	0.3383	0.2867	0.5325 ± 0.0136	0.5804 ± 0.0189	$\mathbf{0.6178} \pm 0.0108$	0.6160 ± 0.0085
	QWK	0.0211	0.2420	0.1853	0.1535	0.3704 ± 0.0877	0.5880 ± 0.0763	$\mathbf{0.6767} \pm 0.0811$	0.6231 ± 0.0871

Table 2: Concept and severity classification results on the english CheXpert dataset. Results are given for the overall concept classification and severity classification over all concepts as well as separate for pulmonary edema, pleural effusion and cardiomegaly.

5.2 Report Retrieval

Quantitative results for the report retrieval on the internal dataset are shown in Tab.3. It can be seen that for the exact matching (threshold 0.0), all LASER configurations outperform the baselines on every metric with the best results for LASER_{200}. For the retrieval with a threshold of 0.5, LASER_{200} still outperforms the baselines for P@5, P@10, NDCG@5 and NDCG@10. Here, the gap between LASER_{200} and the baselines widens as more retrieved reports are considered (@10), indicating that LASER maintains higher precision across a larger neighborhood. This suggests that the structured feature space of LASER organizes reports with similar severity profiles into coherent, dense clusters, rather than placing only the single nearest neighbor correctly while the broader neighborhood remains unstructured. Results for the report retrieval on the CheXpert dataset are shown in Tab.4. Again, all LASER configurations outperform the baselines on the exact retrieval (threshold 0.0). Here, the best performing configuration overall is LASER_{800}. Moreover, all LASER configurations outperform the baselines for the retrieval with a threshold of 0.5. Here, the different configurations achieve almost identical results with the best performance overall for LASER_{1200} and LASER_{800}. As already described for the cardiac dataset, the performance gap between LASER and the baselines widens as more retrieved reports are included in the evaluation. Concept-specific retrieval results for both datasets are shown in Appendix A.2.1. Qualitative examples of the retrieval on the CheXpert dataset are included in Appendix A.2.3.

5.3 Latent Space

To assess whether LASER induces a geometrically structured latent space, we analyze the embedding space of real radiology report sentences. Qualitative results of the structured latent space can be seen in Appendix A.2.2. To quantify the severity structure, we compute the Spearman-correlation coefficient between the first PCA component of each concept cluster and the corresponding severity

Method	P@1	P@5	P@10	NDCG@1	NDCG@5	NDCG@10	MRR
Threshold 0.0							
German BERT	0.3893	0.3389	0.3198	0.3893	0.3509	0.3359	0.4826
German MedBERT	0.3969	0.3328	0.3214	0.3969	0.3464	0.3358	0.4872
RadBERT	0.3664	0.3176	0.2824	0.3664	0.3252	0.3008	0.4493
Multilingual E5-large	0.4046	0.3847	0.3664	0.4046	0.3888	0.3783	0.5088
LASER_{200}	$\mathbf{0.4809} \pm 0.0378$	$\mathbf{0.4293} \pm 0.0128$	$\mathbf{0.4006} \pm 0.0155$	$\mathbf{0.4809} \pm 0.0378$	$\mathbf{0.4410} \pm 0.0166$	$\mathbf{0.4210} \pm 0.0168$	$\mathbf{0.5763} \pm 0.0275$
LASER_{400}	0.4458 ± 0.0289	0.4058 ± 0.0127	0.3838 ± 0.0062	0.4458 ± 0.0289	0.4166 ± 0.0121	0.4036 ± 0.0072	0.5502 ± 0.0146
LASER_{800}	0.4107 ± 0.0303	0.4037 ± 0.0144	0.3799 ± 0.0053	0.4107 ± 0.0303	0.4083 ± 0.0181	0.3966 ± 0.0096	0.5321 ± 0.0224
LASER_{1200}	0.4320 ± 0.0116	0.4205 ± 0.0060	0.3959 ± 0.0045	0.4320 ± 0.0116	0.4240 ± 0.0059	0.4111 ± 0.0047	0.5475 ± 0.0079
Threshold 0.5							
German BERT	0.6718	0.6336	0.6069	0.6718	0.6412	0.6216	0.7654
German MedBERT	0.7023	0.6595	0.6244	0.7023	0.6701	0.6443	0.7973
RadBERT	0.6260	0.5588	0.5382	0.6260	0.5703	0.5549	0.7169
Multilingual E5-large	**0.7405**	0.6824	0.6496	**0.7405**	0.6920	0.6690	**0.8108**
LASER_{200}	0.7206 ± 0.0149	$\mathbf{0.6904} \pm 0.0086$	$\mathbf{0.6763} \pm 0.0088$	0.7206 ± 0.0149	$\mathbf{0.6973} \pm 0.0076$	$\mathbf{0.6875} \pm 0.0063$	0.7948 ± 0.0083
LASER_{400}	0.6809 ± 0.0137	0.6742 ± 0.0111	0.6646 ± 0.0085	0.6809 ± 0.0137	0.6774 ± 0.0078	0.6731 ± 0.0070	0.7738 ± 0.0045
LASER_{800}	0.6779 ± 0.0099	0.6803 ± 0.0111	0.6588 ± 0.0035	0.6779 ± 0.0099	0.6800 ± 0.0102	0.6687 ± 0.0050	0.7735 ± 0.0084
LASER_{1200}	0.6916 ± 0.0308	0.6867 ± 0.0059	0.6722 ± 0.0064	0.6916 ± 0.0308	0.6877 ± 0.0087	0.6810 ± 0.0065	0.7867 ± 0.0208

Table 3: Report retrieval results on the cardiac dataset overall with a relevance threshold of 0.0 and 0.5.

Method	P@1	P@5	P@10	P@20	NDCG@1	NDCG@5	NDCG@10	NDCG@20	MRR
Threshold 0.0									
BERT	0.3711	0.3284	0.3100	0.3006	0.3711	0.3379	0.3215	0.3109	0.5326
BioClinical BERT	0.3840	0.3241	0.3074	0.2948	0.3840	0.3361	0.3203	0.3069	0.5438
RadBERT	0.3883	0.3140	0.3006	0.2971	0.3883	0.3282	0.3138	0.3070	0.5426
Multilingual E5-large	0.4441	0.3782	0.3625	0.3388	0.4441	0.3920	0.3765	0.3549	0.5913
LASER_{200}	0.5095 ± 0.0093	0.4535 ± 0.0068	0.4272 ± 0.0085	0.4006 ± 0.0105	0.5095 ± 0.0093	0.4659 ± 0.0055	0.4433 ± 0.0071	0.4190 ± 0.0091	0.6381 ± 0.0081
LASER_{400}	0.5332 ± 0.0102	0.4905 ± 0.0035	0.4661 ± 0.0039	0.4380 ± 0.0055	0.5332 ± 0.0102	0.4991 ± 0.0037	0.4793 ± 0.0039	0.4550 ± 0.0050	0.6517 ± 0.0054
LASER_{800}	0.5327 ± 0.0105	$\mathbf{0.4960} \pm 0.0066$	$\mathbf{0.4738} \pm 0.0087$	$\mathbf{0.4515} \pm 0.0095$	0.5327 ± 0.0105	$\mathbf{0.5037} \pm 0.0068$	$\mathbf{0.4857} \pm 0.0077$	$\mathbf{0.4659} \pm 0.0083$	$\mathbf{0.6485} \pm 0.0081$
LASER_{1200}	$\mathbf{0.5436} \pm 0.0176$	0.4840 ± 0.0068	0.4620 ± 0.0040	0.4352 ± 0.0117	$\mathbf{0.5436} \pm 0.0176$	0.4960 ± 0.0046	0.4766 ± 0.0021	0.4527 ± 0.0077	0.6517 ± 0.0115
Threshold 0.5									
BERT	0.6934	0.6702	0.6547	0.6380	0.6934	0.6769	0.6633	0.6486	0.8143
BioClinical BERT	0.7278	0.6688	0.6532	0.6366	0.7278	0.6808	0.6657	0.6494	0.8342
RadBERT	0.7307	0.6745	0.6536	0.6413	0.7307	0.6851	0.6667	0.6537	0.8361
Multilingual E5-large	0.7837	0.7301	0.7152	0.6907	0.7837	0.7412	0.7269	0.7055	0.8690
LASER_{200}	0.8255 ± 0.0091	0.7904 ± 0.0066	0.7728 ± 0.0053	0.7521 ± 0.0052	0.8255 ± 0.0091	0.7977 ± 0.0061	0.7830 ± 0.0051	0.7647 ± 0.0046	$\mathbf{0.8893} \pm 0.0052$
LASER_{400}	0.8183 ± 0.0053	0.7986 ± 0.0055	$\mathbf{0.7800} \pm 0.0056$	0.7660 ± 0.0050	0.8183 ± 0.0053	0.8028 ± 0.0040	$\mathbf{0.7887} \pm 0.0046$	0.7756 ± 0.0045	0.8838 ± 0.0028
LASER_{800}	0.8244 ± 0.0159	$\mathbf{0.7996} \pm 0.0073$	0.7782 ± 0.0079	$\mathbf{0.7668} \pm 0.0065$	0.8244 ± 0.0159	$\mathbf{0.8055} \pm 0.0072$	0.7886 ± 0.0068	$\mathbf{0.7767} \pm 0.0060$	0.8854 ± 0.0089
LASER_{1200}	$\mathbf{0.8263} \pm 0.0078$	0.7981 ± 0.0049	0.7711 ± 0.0065	0.7522 ± 0.0083	$\mathbf{0.8263} \pm 0.0078$	0.8050 ± 0.0037	0.7838 ± 0.0051	0.7658 ± 0.0064	0.8892 ± 0.0052

Table 4: Report retrieval results on the CheXpert dataset overall with a relevance threshold of 0.0 and 0.5.

grade. A high correlation indicates that the severity is encoded as a continuous geometric axis within the concept cluster, rather than as unstructured variation. Results for the internal cardiac dataset and CheXpert dataset can be seen in Tab.5. LASER achieves better correlation coefficients, i.e. closer to 1 or -1, compared to the baselines on both datasets for all clinical concepts. For the cardiac dataset LASER_{200} is again the best performing configuration with correlation coefficients of -0.907, -0.605, -0.779 and -0.783 for calcification, stenosis, aorta/tp and myocardium. For the CheXpert dataset, LASER_{800} performs best overall with correlation coefficients of -0.531, -0.630 and -0.701 for the concepts edema, pulmonary effusion and cardiomegaly. It can be seen that LASER is capable of structuring a geometric, severity-aware feature space with synthetic anchors during training.

5.4 Ablation Studies

We conduct two ablation studies with full results in AppendixA.1. Firstly, we assess the contribution of each loss term by removing them individually during training. On both datasets, removing any of the contrastive loss term tends to reduce severity classification performance in terms of accuracy, F1-score, and QWK. In contrast, concept classification remains largely unaffected by the loss ablations, as the concept-specific classification loss is retained in every configuration. Removing $\mathcal{L}_{\text{Concept}}$ has a notable effect on concept classification performance on the German cardiac CT dataset (Accuracy

Method	Cardiac				CheXpert		
	Calcification	Stenosis	Aorta/TP	Myocardium	Edema	Pleural Effusion	Cardiomegaly
(German) BERT	0.141	0.345	-0.357	-0.505	0.028	0.355	-0.209
German MedBERT / BioClinical BERT	-0.620	0.181	0.759	-0.706	-0.018	0.226	-0.144
RadBERT	0.006	0.266	-0.043	0.553	0.010	0.279	-0.068
Multilingual E5-large	0.803	-0.566	0.691	-0.689	0.144	0.479	0.034
LASER_{200}	$\mathbf{0.9046} \pm 0.0017$	$\mathbf{0.5822} \pm 0.0401$	0.7776 ± 0.0013	0.7980 ± 0.0020	0.4398 ± 0.0177	0.6184 ± 0.0446	0.5350 ± 0.1210
LASER_{400}	0.9018 ± 0.0029	0.5318 ± 0.0283	0.7762 ± 0.0026	0.7958 ± 0.0167	0.4942 ± 0.0223	0.6276 ± 0.0067	0.6662 ± 0.0178
LASER_{800}	0.8912 ± 0.0047	0.5626 ± 0.0228	$\mathbf{0.7808} \pm 0.0037$	0.7828 ± 0.0216	$\mathbf{0.5178} \pm 0.0565$	0.6312 ± 0.0260	$\mathbf{0.6838} \pm 0.0094$
LASER_{1200}	0.8966 ± 0.0018	0.5402 ± 0.0078	0.7774 ± 0.0024	$\mathbf{0.8106} \pm 0.0044$	0.4856 ± 0.0301	$\mathbf{0.6452} \pm 0.0092$	0.6494 ± 0.0367

Table 5: Spearman-correlation coefficients between the first PCA component of each concept cluster and the corresponding severity grade for the internal cardiac dataset and the CheXpert dataset.

drops from 0.924 to 0.884), while severity classification remains largely unaffected on both datasets, suggesting that this loss primarily enforces concept separation in the embedding space rather than contributing to severity ordering. Secondly, we evaluate the advantage of ordinal geometric supervision compared to categorical supervision on the same data. Higher QWK values paired with lower standard deviations for our ordinal supervision compared to using the cross-entropy loss on the same data shows that ordinal geometric supervision, not merely exposure to severity labels, is the key factor to ensure a structured severity-aware latent space. On CheXpert, LASER outperforms both classifiers in all metrics for every training set of 400 or more sentences per concept, improving QWK over the Classifier by 0.09–0.13. On the smaller German cardiac dataset there is no clear advantage: the classifier achieves the highest QWK at 200 sentences per concept (0.795 vs. 0.764), while LASER attains higher QWK at 400–1200 sentences. Differences are largely within seed variance. We hypothesize that the stereotyped phrasing of German cardiac reports, where severity is often expressed through a limited set of fixed terms, reduces the benefit of geometric supervision. Consequently, a classifier can reliably separate the discrete severity categories without requiring a continuous ordinal axis in the latent space. The advantage of LASER's geometric supervision is more pronounced on CheXpert, where greater linguistic variation in English chest reports makes categorical boundaries less reliable and ordinal structure more beneficial. Thirdly, we compare the performance of our LASER models to a configuration where the same synthetic data is used during training, but the Rank-N-Contrast (RNC) loss Zha et al. (2023) is utilized. The results again show consistently higher QWK values for our LASER model compared to the RNC trained model, suggesting that our multi-objective loss induces a more structured severity-aware feature space.

6 Discussion

We propose LASER, a label-efficient framework to train a concept- and severity-aware text encoder with a structured latent space. Unlike generic text encoders such as BERT, MedBERT, BioClinical BERT, RadBERT, and even the strong multilingual E5-large, where concept and severity distinctions collapse in the latent space due to the marginal representation of specialized clinical text during pre-training, our multi-objective training explicitly separates clinical concepts and imposes a continuous severity ordering within each concept cluster. This structured geometry is achieved entirely without manual annotation, as an LLM generates synthetic sentences for each concept–severity combination at zero labeling cost, providing a structured training signal that directly shapes the target feature space.

LASER consistently outperforms state-of-the-art text encoders on two datasets across different languages and organ systems on both severity classification and report retrieval. This reflects a general principle: explicitly supervising the geometry of the latent space with concept and severity structure yields representations that are both more discriminative and more retrievable than those produced by generic pre-training objectives. The geometric severity structure is further quantified by the Spearman-correlation coefficient between the first PCA component and severity grade, which is higher for LASER than for all baselines, confirming that ordinal geometric supervision induces a continuous severity axis in the latent space that is absent in off-the-shelf encoders.

Ablation studies confirm that each loss component contributes to the ordinal quality of the learned representations, with the full multi-objective objective consistently achieving the highest QWK. Notably, replacing our loss with the Rank-N-Contrast loss Zha et al. (2023) yields lower QWK on both datasets, demonstrating that our specific combination of concept separation and severity ordering losses is more effective. Furthermore, replacing our ordinal geometric supervision with categorical cross-entropy training on the same synthetic sentences consistently yields lower QWK and higher variance across runs, confirming that geometric ordinal supervision is the key factor enabling LASER's structured latent space.

A key advantage of LASER is its interpretability by design. Since each region of the feature space corresponds to a known clinical concept at a defined severity level, the position of a sentence embedding directly encodes its clinical meaning. This enables both severity classification and report retrieval to be performed via simple nearest-neighbor search in the structured space, without any task-specific fine-tuning on real-world data.

Beyond the specific tasks evaluated here, LASER's label-efficient paradigm is transferable to other clinical domains where severity-graded free-text findings are documented, making it a broadly applicable framework. The structured latent space could further serve as a foundation for downstream tasks such as longitudinal disease progression monitoring or severity-aware report generation.

AI USE STATEMENT

In this work, we used generative AI tools for [Generate synthetic data sets, design or provide feedback on research methodology or experiments]. We have not used generative AI tools for [implement methods, help develop theoretical models or conceptual frameworks, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results], and [formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses] are not applicable to this work. Additionally, we used generative AI tools for [create or edit software code, summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, propose a title or keywords for a research paper]. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our results. Full implementation details, including training hyperparameters, optimizer settings, learning rate schedule, and the random seeds used across all experiments and ablation studies are provided in Appendix A.4.1 and Table 18. The prompts used to generate synthetic sentences are provided in AppendixA.4.2. The full code for training and evaluation as well as the synthetic sentences utilized during training will be made publicly available upon acceptance. The CheXpert dataset used in our experiments is publicly available.

ACKNOWLEDGMENTS

REFERENCES

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Anthropic. Claude (claude sonnet 5). `https://www.anthropic.com`, 2026. Large Language Model.

Keno K Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P Loyen, Stefan M Niehues, et al. medbert. de: a comprehensive german bert model for the medical domain. *Expert Systems with Applications*, 237:121598, 2024.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. In *Proceedings of the 28th international conference on computational linguistics*, pp. 6788–6796, 2020.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024.

Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *International Conference on Learning Representations*, 2019.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Wenqiang Tang, Zhouwang Yang, and Yanzhi Song. Disease-grading networks with ordinal regularization for medical imaging. *Neurocomputing*, 545:126245, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024a.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 11897–11916, 2024b.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36:17882–17903, 2023.

Tengfei Zhang, Ziheng Zhao, Chaoyi Wu, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radir: a scalable framework for multi-grained medical image retrieval via radiology report mining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 508–518. Springer, 2025.

Aoxiao Zhong, Xiang Li, Dufan Wu, Hui Ren, Kyungsang Kim, Younggon Kim, Varun Buch, Nir Neumark, Bernardo Bizzo, Won Young Tak, et al. Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in covid-19. *Medical Image Analysis*, 70: 101993, 2021.

A Appendix

A.1 Ablation Study

A.1.1 Loss Functions

To evaluate the influence of the five loss functions on the results, we evaluated the severity and concept classification on both datasets when excluding individual loss terms during training. For this, we use the best-performing LASER sentence-length configuration for each dataset, i.e. 200 sentences for the cardiac dataset and 800 sentences for the CheXpert dataset. All results are averaged over 5 random seeds (mean ± standard deviation). The results for the cardiac dataset can be seen in Tab.6 and for the CheXpert dataset in Tab.7. In both cases, removing any of the severity controlling loss term tends to reduce severity classification performance in terms of accuracy, F1-score, and QWK, although the differences for most terms lie within the range of seed-to-seed variance. Overall, the performance drops are most pronounced for the cosine regression loss and the multiple-negative-ranking loss, compared to the triplet and hard-negative loss. In contrast, concept classification remains largely unaffected by the loss ablations, as the concept-specific cross-entropy loss is retained in every configuration. When removing the concept cross-entropy loss, the severity prediction performance does not drop, showing it has no direct influence on the severity-ordering of the feature

space. However, removing the concept loss has a notable effect on the concept classification performance on the German cardiac CT dataset (Accuracy drops from 0.924 to 0.884). This suggests, that $\mathcal{L}_{\text{Concept}}$ primarily enforces concept separation in the embedding space rather than contributing to severity ordering. On the CheXpert dataset, the effect on concept classification is negligible, likely because the multilingual E5-large encoder already provides strong concept-discriminative representations for english clinical text, reducing the need for explicit concept supervision. Nevertheless, we retain $\mathcal{L}_{\text{Concept}}$ in the full LASER objective as it provides a consistent and interpretable concept structure across datasets and languages.

Configuration	Concept		Severity		
	Acc	F1	Acc	F1	QWK
Full LASER	0.9236 ± 0.0207	0.9268 ± 0.0193	0.8250 ± 0.0164	0.6971 ± 0.0174	0.7640 ± 0.0636
-Cosine Regression loss	0.9117 ± 0.0239	0.9147 ± 0.0225	0.8172 ± 0.0044	0.6962 ± 0.0062	0.7197 ± 0.0091
-Hard-Negative loss	0.9283 ± 0.0168	0.9312 ± 0.0159	0.8171 ± 0.0130	0.6914 ± 0.0139	0.7632 ± 0.0595
-Triplet loss	0.9283 ± 0.0172	0.9308 ± 0.0162	0.8159 ± 0.0129	0.6896 ± 0.0168	0.7326 ± 0.0337
-Multiple-Negative-Ranking loss	0.9137 ± 0.0079	0.9157 ± 0.0078	0.8167 ± 0.0107	0.6829 ± 0.0207	0.7118 ± 0.0214
-Concept loss	0.8841 ± 0.0066	0.8895 ± 0.0064	0.8283 ± 0.0103	0.7059 ± 0.0082	0.7646 ± 0.0469

Table 6: Concept and severity classification results on the german cardiac CT dataset when excluding single loss functions compared to the full training set-up.

Configuration	Concept		Severity		
	Acc	F1	Acc	F1	QWK
Full LASER	0.9861 ± 0.0018	0.9857 ± 0.0018	0.8022 ± 0.0121	0.6854 ± 0.0347	0.6324 ± 0.0376
-Cosine Regression loss	0.9863 ± 0.0026	0.9860 ± 0.0026	0.7954 ± 0.0281	0.6692 ± 0.0413	0.6195 ± 0.0570
-Hard-Negative loss	0.9863 ± 0.0018	0.9860 ± 0.0019	0.7918 ± 0.0246	0.6631 ± 0.0408	0.6090 ± 0.0560
-Triplet loss	0.9866 ± 0.0018	0.9862 ± 0.0019	0.7978 ± 0.0094	0.6789 ± 0.0191	0.6292 ± 0.0217
-Multiple-Negative-Ranking loss	0.9823 ± 0.0013	0.9815 ± 0.0014	0.7670 ± 0.0270	0.6360 ± 0.0483	0.5488 ± 0.0749
-Concept loss	0.9862 ± 0.0015	0.9858 ± 0.0015	0.7944 ± 0.0043	0.7033 ± 0.0118	0.6367 ± 0.0097

Table 7: Concept and severity classification results on the CheXpert dataset when excluding single loss functions compared to the full training set-up.

A.1.2 Categorical Classification

To evaluate the advantage of using ordinal geometric supervision with the synthetic sentences compared to categorical supervision with the same data, we fine-tune the same text encoder with a cross-entropy loss on the severity label and compare the results. Severity classification results for the internal cardiac dataset are shown in Tab.8. While the Classifier achieves marginally higher accuracy in some configurations, LASER consistently outperforms it on QWK across all sentence counts except 200. This metric directly quantifies the ordinal quality of the learned representations. For LASER_{400}, LASER_{800}, and LASER_{1200}, the QWK improvement over the categorical baseline is substantial, while the standard deviation is consistently lower, indicating that ordinal geometric supervision not only improves performance but also produces more stable and consistent latent representations across runs. The accuracy advantage of the Classifier in some configurations is therefore misleading as it reflects correct classification of individual grades without capturing their ordinal relationships, as evidenced by the lower QWK. The same is true for the CheXpert dataset, as shown in Tab.9, where all configurations, except 200, achieve lower QWK when trained with cross-entropy supervision compared to their ordinally supervised counterparts. Furthermore, cross-entropy supervision again yields substantially higher standard deviations across runs, suggesting the same instability as on the cardiac dataset. These results demonstrate that ordinal geometric supervision, not merely exposure to severity labels, is the key factor to ensure a structured latent space.

A.1.3 Rank-N-Contrast Loss

To evaluate our multi-objective loss, we also fine-tune the same text encoder with the same synthetic sentences using the Rank-N-Contrast (RNC) loss Zha et al. (2023) instead of our multi-objective ordinal loss. The results on the cardiac dataset can be seen in Tab.10 and for the CheXpert dataset in Tab.11. On the CheXpert dataset, LASER outperforms RNC across all sentence configurations on all three metrics, with particularly large margins in QWK (e.g., 0.606 vs. 0.368 for 400 sentences).

Configuration	Acc	F1	QWK
Classifier_{200}	$\mathbf{0.8329} \pm 0.0091$	$\mathbf{0.7100} \pm 0.0187$	$\mathbf{0.7946} \pm 0.0339$
LASER_{200}	0.8250 ± 0.0164	0.6971 ± 0.0174	0.7640 ± 0.0636
Classifier_{400}	$\mathbf{0.7916} \pm 0.0187$	0.6689 ± 0.0213	0.6967 ± 0.0776
LASER_{400}	0.7854 ± 0.0210	$\mathbf{0.6794} \pm 0.0234$	$\mathbf{0.7390} \pm 0.0532$
Classifier_{800}	$\mathbf{0.7773} \pm 0.0350$	0.6548 ± 0.0272	0.6737 ± 0.0829
LASER_{800}	0.7633 ± 0.0134	$\mathbf{0.6621} \pm 0.0200$	$\mathbf{0.7100} \pm 0.0488$
Classifier_{1200}	$\mathbf{0.7581} \pm 0.0252$	0.6341 ± 0.0242	0.6319 ± 0.0479
LASER_{1200}	0.7572 ± 0.0094	$\mathbf{0.6553} \pm 0.0093$	$\mathbf{0.6853} \pm 0.0308$

Table 8: Severity classification results on the German cardiac CT dataset comparing LASER against a categorical cross-entropy baseline (Classifier) trained on the same synthetic sentences.

Configuration	Acc	F1	QWK
Classifier_{200}	$\mathbf{0.7383} \pm 0.0286$	0.5745 ± 0.0460	0.4528 ± 0.0617
LASER_{200}	0.7079 ± 0.0234	$\mathbf{0.5947} \pm 0.0369$	$\mathbf{0.4586} \pm 0.0808$
Classifier_{400}	0.7536 ± 0.0622	0.5941 ± 0.0916	0.5184 ± 0.1164
LASER_{400}	$\mathbf{0.7785} \pm 0.0134$	$\mathbf{0.6812} \pm 0.0259$	$\mathbf{0.6059} \pm 0.0318$
Classifier_{800}	0.7614 ± 0.0428	0.5942 ± 0.0770	0.5278 ± 0.0874
LASER_{800}	$\mathbf{0.8022} \pm 0.0121$	$\mathbf{0.6854} \pm 0.0347$	$\mathbf{0.6324} \pm 0.0376$
Classifier_{1200}	0.7122 ± 0.0897	0.5517 ± 0.1131	0.4531 ± 0.1447
LASER_{1200}	$\mathbf{0.7909} \pm 0.0082$	$\mathbf{0.6446} \pm 0.0208$	$\mathbf{0.5816} \pm 0.0286$

Table 9: Severity classification results on the CheXpert dataset comparing LASER against a categorical cross-entropy baseline (Classifier) trained on the same synthetic sentences.

On the cardiac dataset, RNC achieves competitive or higher accuracy and F1 in some configurations, but LASER consistently outperforms RNC on QWK across all sentence counts. Again, as this metric directly quantifies the ordinal quality of the learned representations, penalizing predictions proportionally to their distance from the true grade, these results demonstrate that LASER's multi-objective loss induces a more structured severity-aware feature space than a single ordinal contrastive loss, despite the latter being specifically designed for regression tasks with continuous target ordering.

Configuration	Acc	F1	QWK
$\text{Rank-N-Contrast}_{200}$	0.7938 ± 0.0070	0.6564 ± 0.0106	0.6972 ± 0.0220
LASER_{200}	$\mathbf{0.8250} \pm 0.0164$	$\mathbf{0.6971} \pm 0.0174$	$\mathbf{0.7640} \pm 0.0636$
$\text{Rank-N-Contrast}_{400}$	$\mathbf{0.7989} \pm 0.0087$	0.6364 ± 0.0106	0.6882 ± 0.0236
LASER_{400}	0.7854 ± 0.0210	$\mathbf{0.6794} \pm 0.0234$	$\mathbf{0.7390} \pm 0.0532$
$\text{Rank-N-Contrast}_{800}$	$\mathbf{0.7989} \pm 0.0121$	$\mathbf{0.7187} \pm 0.0143$	0.6964 ± 0.0215
LASER_{800}	0.7633 ± 0.0134	0.6621 ± 0.0200	$\mathbf{0.7100} \pm 0.0488$
$\text{Rank-N-Contrast}_{1200}$	$\mathbf{0.8816} \pm 0.0077$	$\mathbf{0.6639} \pm 0.0087$	0.6602 ± 0.0142
LASER_{1200}	0.7572 ± 0.0094	0.6553 ± 0.0093	$\mathbf{0.6853} \pm 0.0308$

Table 10: Severity classification results on the German cardiac CT dataset comparing LASER against a Rank-N-Contrast baseline trained on the same synthetic sentences.

A.2 Additional Results

A.2.1 Concept Specific Retrieval Results

Concept-specific report retrieval results on the cardiac dataset are shown in Tab.12. We can see that LASER outperforms the baselines for the concepts calcification, myocardium and aorta/tp.

Configuration	Acc	F1	QWK
Rank-N-Contrast$_{200}$	0.6062 ± 0.0234	0.4344 ± 0.0237	0.3394 ± 0.0063
LASER$_{200}$	$\mathbf{0.7079} \pm 0.0234$	$\mathbf{0.5947} \pm 0.0369$	$\mathbf{0.4586} \pm 0.0808$
Rank-N-Contrast$_{400}$	0.6848 ± 0.0333	0.5253 ± 0.0411	0.3677 ± 0.0128
LASER$_{400}$	$\mathbf{0.7785} \pm 0.0134$	$\mathbf{0.6812} \pm 0.0259$	$\mathbf{0.6059} \pm 0.0318$
Rank-N-Contrast$_{800}$	0.7324 ± 0.0104	0.5748 ± 0.0099	0.4120 ± 0.0298
LASER$_{800}$	$\mathbf{0.8022} \pm 0.0121$	$\mathbf{0.6854} \pm 0.0347$	$\mathbf{0.6324} \pm 0.0376$
Rank-N-Contrast$_{1200}$	0.7428 ± 0.0043	0.5792 ± 0.0060	0.4528 ± 0.0065
LASER$_{1200}$	$\mathbf{0.7909} \pm 0.0082$	$\mathbf{0.6446} \pm 0.0208$	$\mathbf{0.5816} \pm 0.0286$

Table 11: Severity classification results on the CheXpert dataset comparing LASER against a Rank-N-Contrast baseline trained on the same synthetic sentences.

Method	P@1	P@5	P@10	NDCG@1	NDCG@5	NDCG@10	MRR
Calcification							
German BERT	0.7297	0.6505	0.6333	0.7297	0.6704	0.6521	0.8001
German MedBERT	0.7568	0.7009	0.6586	0.7568	0.7096	0.6774	0.8303
RadBERT	0.6306	0.5766	0.5559	0.6306	0.5857	0.5679	0.7163
Multilingual E5-large	0.7477	0.7369	0.7153	0.7477	0.7399	0.7243	0.8402
LASER$_{200}$ sentence	$\mathbf{0.8468} \pm 0.0211$	0.8436 ± 0.0061	$\mathbf{0.8400} \pm 0.0046$	$\mathbf{0.8468} \pm 0.0211$	0.8446 ± 0.0048	$\mathbf{0.8420} \pm 0.0044$	$\mathbf{0.8992} \pm 0.0127$
LASER$_{400}$ sentence	0.8288 ± 0.0229	0.8230 ± 0.0087	0.8266 ± 0.0044	0.8288 ± 0.0229	0.8248 ± 0.0099	0.8267 ± 0.0054	0.8836 ± 0.0148
LASER$_{800}$ sentence	0.8378 ± 0.0191	0.8220 ± 0.0175	0.8090 ± 0.0228	0.8378 ± 0.0191	0.8241 ± 0.0168	0.8144 ± 0.0206	0.8816 ± 0.0137
LASER$_{1200}$ sentence	0.8450 ± 0.0118	$\mathbf{0.8472} \pm 0.0099$	0.8211 ± 0.0149	0.8450 ± 0.0118	$\mathbf{0.8462} \pm 0.0103$	0.8285 ± 0.0138	0.8933 ± 0.0093
Stenosis							
German BERT	0.8390	0.8186	0.8042	0.8390	0.8245	0.8166	0.8773
German MedBERT	0.8305	0.8254	0.8102	0.8305	0.8278	0.8189	0.8723
RadBERT	0.7966	0.7847	0.7814	0.7966	0.7893	0.7875	0.8516
Multilingual E5-large	0.8644	**0.8593**	**0.8466**	0.8644	**0.8619**	**0.8567**	0.9057
LASER$_{200}$ sentence	$\mathbf{0.8763} \pm 0.0176$	0.8546 ± 0.0065	0.8351 ± 0.0112	$\mathbf{0.8763} \pm 0.0176$	0.8604 ± 0.0090	0.8559 ± 0.0107	$\mathbf{0.9111} \pm 0.0118$
LASER$_{400}$ sentence	0.8152 ± 0.0219	0.8173 ± 0.0215	0.8041 ± 0.0218	0.8152 ± 0.0219	0.8188 ± 0.0198	0.8176 ± 0.0220	0.8726 ± 0.0141
LASER$_{800}$ sentence	0.7864 ± 0.0220	0.7983 ± 0.0056	0.7880 ± 0.0111	0.7864 ± 0.0220	0.7980 ± 0.0095	0.7976 ± 0.0120	0.8551 ± 0.0154
LASER$_{1200}$ sentence	0.8220 ± 0.0344	0.7936 ± 0.0088	0.7786 ± 0.0148	0.8220 ± 0.0344	0.8000 ± 0.0109	0.7926 ± 0.0154	0.8713 ± 0.0191
Myocardium							
German BERT	0.7476	0.6951	0.6767	0.7476	0.7114	0.7005	0.8110
German MedBERT	0.7670	0.7417	0.7272	0.7670	0.7577	0.7475	0.8343
RadBERT	0.7864	0.7087	0.6942	0.7864	0.7292	0.7155	0.8396
Multilingual E5-large	0.8350	0.7573	0.7437	0.8350	0.7746	0.7656	0.8747
LASER$_{200}$ sentence	0.8835 ± 0.0137	0.8579 ± 0.0116	0.8227 ± 0.0141	0.8835 ± 0.0137	0.8695 ± 0.0098	0.8477 ± 0.0113	0.9186 ± 0.0085
LASER$_{400}$ sentence	0.8816 ± 0.0260	0.8551 ± 0.0099	0.8181 ± 0.0233	0.8816 ± 0.0260	0.8653 ± 0.0094	0.8429 ± 0.0178	0.9166 ± 0.0135
LASER$_{800}$ sentence	0.8816 ± 0.0187	0.8691 ± 0.0154	0.8167 ± 0.0183	0.8816 ± 0.0187	0.8797 ± 0.0139	0.8459 ± 0.0139	0.9156 ± 0.0129
LASER$_{1200}$ sentence	$\mathbf{0.8854} \pm 0.0106$	$\mathbf{0.8742} \pm 0.0124$	$\mathbf{0.8501} \pm 0.0149$	$\mathbf{0.8854} \pm 0.0106$	$\mathbf{0.8833} \pm 0.0107$	$\mathbf{0.8693} \pm 0.0127$	$\mathbf{0.9182} \pm 0.0072$
Aorta/TP							
German BERT	0.8644	0.7712	0.7534	0.8644	0.7898	0.7783	0.8905
German MedBERT	0.8814	0.8356	0.8161	0.8814	0.8421	0.8334	0.9142
RadBERT	0.8051	0.7746	0.7610	0.8051	0.7850	0.7802	0.8678
Multilingual E5-large	0.8559	0.8169	0.8051	0.8559	0.8257	0.8225	0.8970
LASER$_{200}$ sentence	0.8898 ± 0.0060	$\mathbf{0.8793} \pm 0.0074$	$\mathbf{0.8597} \pm 0.0063$	0.8898 ± 0.0060	$\mathbf{0.8835} \pm 0.0060$	$\mathbf{0.8861} \pm 0.0063$	0.9260 ± 0.0036
LASER$_{400}$ sentence	0.8814 ± 0.0085	0.8739 ± 0.0039	0.8576 ± 0.0035	0.8814 ± 0.0085	0.8780 ± 0.0061	0.8844 ± 0.0023	0.9222 ± 0.0065
LASER$_{800}$ sentence	0.8848 ± 0.0096	0.8620 ± 0.0154	0.8497 ± 0.0052	0.8848 ± 0.0096	0.8692 ± 0.0147	0.8744 ± 0.0087	0.9236 ± 0.0051
LASER$_{1200}$ sentence	$\mathbf{0.8949} \pm 0.0142$	0.8654 ± 0.0074	0.8414 ± 0.0086	$\mathbf{0.8949} \pm 0.0142$	0.8752 ± 0.0073	0.8732 ± 0.0072	$\mathbf{0.9266} \pm 0.0105$

Table 12: Report retrieval results on the internal cardiac dataset for each concept.

For calcification and myocardium, the best performing configuration is LASER$_{200}$. For aorta/tp all configurations are close to each other with the highest results for LASER$_{800}$ and LASER$_{1200}$. For the concept stenosis, the baseline Multilingual E5-large slightly outperforms LASER$_{200}$ for P@5, P@10, NDCG@5 and NDCG@10. However, this difference is only around 1%, making LASER still the best performing model overall. Tab.13 shows concept-specific retrieval results on the CheXpert dataset. The LASER configurations outperform the baselines on all concepts. Again, the performance gap widens the more retrieved reports are evaluated. For all concepts, the LASER configurations are relatively close to each other with only slight deviations.

A.2.2 Qualitative Structured Latent Space

Fig. 2 and Fig.3 show the t-SNE visualizations of sentence embeddings for the internal dataset and the CheXpert dataset sentences for the baseline text encoder and our LASER text encoder. Figure 3 does the same for the sentences from the CheXpert dataset. Both figures highlight the structured latent space of our LASER model compared to the unstructured baselines, making sentence positions in the latent directly interpretable.

Method	P@1	P@5	P@10	P@20	NDCG@1	NDCG@5	NDCG@10	NDCG@20	MRR
Pulmonary Edema									
BERT	0.6151	0.5502	0.5279	0.5037	0.6151	0.5620	0.5423	0.5206	0.7368
BioClinical BERT	0.7189	0.6423	0.6242	0.5559	0.7189	0.6497	0.6346	0.5829	0.8016
RadBERT	0.7075	0.6551	0.6375	0.5871	0.7075	0.6619	0.6485	0.6098	0.7992
Multilingual E5-large	0.7226	0.6992	0.6709	0.6362	0.7226	0.7054	0.6840	0.6560	0.8230
LASER_{200}	**0.7672** ± 0.0086	0.7170 ± 0.0041	0.6974 ± 0.0032	0.6779 ± 0.0052	**0.7672** ± 0.0086	0.7272 ± 0.0037	0.7118 ± 0.0026	0.6948 ± 0.0034	0.8447 ± 0.0071
LASER_{400}	0.7653 ± 0.0083	**0.7291** ± 0.0101	**0.7123** ± 0.0071	0.6904 ± 0.0099	0.7653 ± 0.0083	**0.7382** ± 0.0074	**0.7255** ± 0.0055	0.7076 ± 0.0075	0.8470 ± 0.0044
LASER_{800}	0.7581 ± 0.0155	0.7253 ± 0.0051	0.7094 ± 0.0060	**0.6942** ± 0.0021	0.7581 ± 0.0155	0.7331 ± 0.0057	0.7209 ± 0.0061	**0.7079** ± 0.0035	**0.8450** ± 0.0088
LASER_{1200}	0.7509 ± 0.0081	0.7164 ± 0.0023	0.7041 ± 0.0100	0.6912 ± 0.0087	0.7509 ± 0.0081	0.7244 ± 0.0037	0.7143 ± 0.0063	0.7035 ± 0.0068	0.8399 ± 0.0053
Pleural Effusion									
BERT	0.6016	0.5503	0.5566	0.5343	0.6016	0.5631	0.5630	0.5444	0.7320
BioClinical BERT	0.6935	0.6371	0.6161	0.5824	0.6935	0.6482	0.6300	0.6017	0.8000
RadBERT	0.7435	0.6919	0.6689	0.6284	0.7435	0.7025	0.6826	0.6492	0.8327
Multilingual E5-large	0.7726	0.7319	0.7100	0.6785	0.7726	0.7423	0.7233	0.6965	0.8553
LASER_{200} sentence	0.7952 ± 0.0077	**0.7770** ± 0.0066	0.7686 ± 0.0068	0.7491 ± 0.0052	0.7952 ± 0.0077	0.7823 ± 0.0048	0.7756 ± 0.0047	0.7607 ± 0.0041	0.8619 ± 0.0023
LASER_{400} sentence	0.7926 ± 0.0104	0.7728 ± 0.0054	0.7646 ± 0.0021	0.7477 ± 0.0022	0.7926 ± 0.0104	0.7787 ± 0.0050	0.7722 ± 0.0029	0.7588 ± 0.0021	0.8604 ± 0.0064
LASER_{800} sentence	0.7997 ± 0.0064	0.7708 ± 0.0069	0.7639 ± 0.0100	0.7490 ± 0.0098	0.7997 ± 0.0064	0.7770 ± 0.0071	0.7705 ± 0.0093	0.7582 ± 0.0095	0.8662 ± 0.0073
LASER_{1200} sentence	**0.8142** ± 0.0207	0.7768 ± 0.0079	**0.7704** ± 0.0081	**0.7586** ± 0.0101	**0.8142** ± 0.0207	**0.78349** ± 0.0100	**0.7767** ± 0.0091	**0.7664** ± 0.0097	**0.8726** ± 0.0129
Cardiomegaly									
BERT	0.3838	0.3725	0.3717	0.3627	0.3838	0.3629	0.3670	0.3637	0.5553
BioClinical BERT	0.8020	0.7228	0.6800	0.6461	0.8020	0.7409	0.7047	0.6722	0.8646
RadBERT	0.8323	0.7523	0.7010	0.6569	0.8323	0.7697	0.7274	0.6867	0.8845
Multilingual E5-large	0.8646	0.8141	0.7636	0.7429	0.8646	0.8279	0.7871	0.7642	0.9113
LASER_{200} sentence	0.9010 ± 0.0029	**0.8725** ± 0.0032	**0.8570** ± 0.0042	**0.8362** ± 0.0024	0.9010 ± 0.0029	**0.8786** ± 0.0029	**0.8656** ± 0.0035	**0.8480** ± 0.0022	**0.9325** ± 0.0024
LASER_{400} sentence	**0.9034** ± 0.0030	0.8685 ± 0.0005	0.8509 ± 0.0048	0.8349 ± 0.0029	**0.9034** ± 0.0030	0.8753 ± 0.0010	0.8609 ± 0.0029	0.8460 ± 0.0021	0.9307 ± 0.0027
LASER_{800} sentence	0.8881 ± 0.0145	0.8619 ± 0.0046	0.8447 ± 0.0050	0.8330 ± 0.0035	0.8881 ± 0.0145	0.8666 ± 0.0057	0.8530 ± 0.0055	0.8418 ± 0.0043	0.9218 ± 0.0091
LASER_{1200} sentence	0.8893 ± 0.0046	0.8634 ± 0.0064	0.8477 ± 0.0068	0.8333 ± 0.0060	0.8893 ± 0.0046	0.8683 ± 0.0061	0.8556 ± 0.0065	0.8428 ± 0.0057	0.9226 ± 0.0030

Table 13: Report retrieval results on the CheXpert dataset for all concepts.

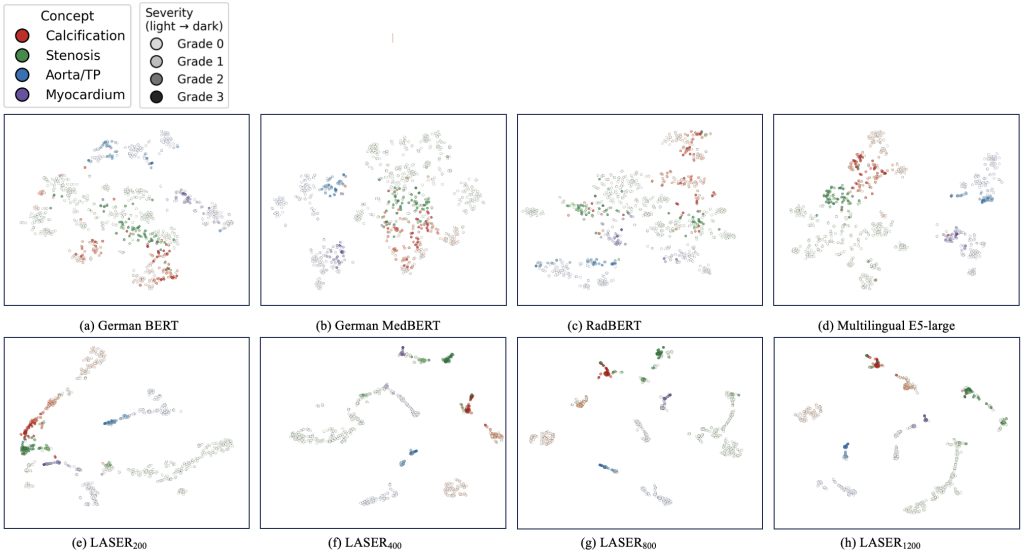


Figure 2: Latent space sentence representations on the cardiac dataset for the baselines BERT, MedBERT, RadBERT and Multilingual E5-large as well as our model LASER with different numbers of synthetic sentences used for training.

A.2.3 Qualitative Report Retrieval

For the best performing model on the CheXpert dataset, i.e. LASER_{800}, qualitative results were also evaluated. Tab.14shows five report examples and the top three retrieved reports for the CheXpert dataset.

Qualitative results on the CheXpert dataset for the concept specific report retrieval were also evaluated. Tab.15 shows five qualitative examples for the concept pulmonary edema, Tab.16 for pleural effusion and Tab.17 for cardiomegaly.

A.3 Additional Figures

A.3.1 Confusion Matrices

Figure 4 shows the confusion matrices for severity classification per concept on the internal dataset, complementing the aggregated results in Tab.1 of the main text. Figure 5 shows the confusion

Query	Top-1 Retrieved	Top-2 Retrieved	Top-3 Retrieved
SMALL BILATERAL PLEURAL EFFUSIONS. MILD PULMONARY EDEMA. 2.MILD CARDIOMEGALY.	MILD CARDIOMEGALY. MILD PULMONARY EDEMA. SMALL BILATERAL PLEURAL EFFUSIONS, LEFT GREATER THAN RIGHT.	SMALL PLEURAL EFFUSIONS ARE SEEN BILATERALLY. PERSISTENT MILD PULMONARY EDEMA. MILD CARDIOMEGALY.	MILD PULMONARY VASCULAR CONGESTION, LIKELY MILD PULMONARY EDEMA. SMALL BILATERAL PLEURAL EFFUSIONS. MILD CARDIOMEGALY.
APPEARANCE OF INCREASED CARDIOMEGALY, THOUGH THIS MAY BE BECAUSE OF SLIGHT LEFTWARD ROTATION. INCREASED INDISTINCTNESS OF VASCULAR MARKINGS, CONSISTENT WITH PULMONARY EDEMA.	Mild cardiomegaly. Interstitial prominence which may be due to edema or atypical infection.	INCREASED INTERSTITIAL MARKINGS WITH SUBPLEURAL RETICULAR OPACITIES, CONSISTENT WITH MILD TO MODERATE PULMONARY EDEMA. STABLE CARDIOMEGALY.	1.INCREASING INDISTINCTNESS OF THE PULMONARY VASCULARITY, POSSIBLY MILD EDEMA. 3.CARDIOMEGALY WITH PERICARDIAC CALCIFICATION, CONSISTENT WITH THE PATIENT'S HISTORY OF CALCIFIC PERICARDITIS.
STABLE CARDIOMEDIASTINAL SILHOUETTE WITH REDEMONSTRATION OF CARDIOMEGALY, TORTUOUS AORTA, AND PROMINENT SUPERIOR MEDIASTINUM. INTERVAL DECREASE IN PULMONARY EDEMA.	Redemonstration of marked cardiomegaly. Mild edema.	SEVERE CARDIOMEGALY. MILD PULMONARY EDEMA.	3.POSSIBLE MINIMAL INCREASE IN PULMONARY EDEMA. 4.CARDIOMEGALY IS GROSSLY STABLE.
MODERATE BILATERAL PLEURAL EFFUSIONS ARE AGAIN SEEN. STABLE CARDIOMEGALY. PERSISTENT MILD PULMONARY INTERSTITIAL EDEMA.	INTERVAL INCREASED RIGHT PLEURAL EFFUSION, NOW MODERATE. SLIGHTLY DECREASED MILD TO MODERATE PULMONARY EDEMA. STABLE CARDIOMEGALY.	STABLE MILD PULMONARY EDEMA. STABLE BILATERAL PLEURAL EFFUSIONS AND BIBASILAR CONSOLIDATION. STABLE CARDIOMEGALY.	INTERVAL INCREASE IN HOMOGENEOUS OPACITY THROUGHOUT BOTH LUNG FIELDS POSSIBLY REPRESENTING PLEURAL EFFUSIONS. MILD PULMONARY EDEMA. STABLE CARDIOMEGALY.
3.SMALL RIGHT PLEURAL EFFUSION IS AGAIN SEEN.', '5.MILD CARDIOMEGALY PERSISTS.	SMALL LEFT PLEURAL EFFUSION IS MILDLY INCREASED. MINOR CARDIOMEGALY MAY BE PRESENT.	2.SMALL BILATERAL PLEURAL EFFUSIONS ARE NOTED. 4.MILD CARDIOMEGALY PERSISTS.	TINY BILATERAL PLEURAL EFFUSIONS MAY BE PRESENT. 4.MINOR CARDIOMEGALY IS SEEN.

Table 14: Qualitative retrieval examples for the CheXpert dataset. For each query report, we show the top-3 retrieved reports based on cosine similarity of the overall embedding.

Query	Top-1 Retrieved	Top-2 Retrieved	Top-3 Retrieved
INCREASED INDISTINCTNESS OF VASCULAR MARKINGS, CONSISTENT WITH PULMONARY EDEMA.	COMPARED TO THE PRIOR EXAMINATION THERE IS INCREASED MODERATE PULMONARY EDEMA WITH AREAS OF CONSOLIDATION.	2.PULMONARY VASCULATURE INDISTINCTNESS LIKELY REFLECTS INCREASED PULMONARY EDEMA.	THERE IS INTERVAL INCREASE IN RETICULAR PROMINENCE, MOST LIKELY REPRESENTING INTERSTITIAL PULMONARY EDEMA.
INTERVAL DECREASE IN PULMONARY EDEMA.	INTERVAL DECREASE IN PULMONARY EDEMA.	INTERVAL DECREASE IN PULMONARY EDEMA.	EVIDENCE OF UNDERLYING CHRONIC FIBROTIC LUNG DISEASE WITH POSSIBLE SUPERIMPOSED MILD PULMONARY EDEMA.
4.MILD EDEMA.	STABLE MILD EDEMA.	STABLE MILD EDEMA.	MINIMAL RESIDUAL EDEMA.
PERSISTENT PULMONARY VASCULAR ENGORGEMENT REFLECTING UNCHANGED MODERATE PULMONARY EDEMA.	2.ACCENTUATED CENTRAL PULMONARY VASCULATURE WITH SUPERIMPOSITION OF RETICULAR PARENCHYMAL MARKINGS, POSSIBLY PULMONARY EDEMA.	PERSISTENT MODERATE PULMONARY EDEMA IS UNCHANGED.	DIFFUSE PROMINENCE OF THE PULMONARY VASCULATURE WITH EFFACEMENT COMPATIBLE WITH PULMONARY EDEMA THAT IS UNCHANGED FROM THE PRIOR EXAMINATION.
PERSISTENT MILD PULMONARY INTERSTITIAL EDEMA.	STABLE, MILD INTERSTITIAL PULMONARY EDEMA.	PERSISTENT MILD-TO-MODERATE PULMONARY EDEMA.	PERSISTENT MILD PULMONARY EDEMA.

Table 15: Qualitative retrieval examples for the CheXpert dataset on the clinical concept pulmonary edema. For each query report, we show the top-3 retrieved reports based on cosine similarity of the overall embedding.

Query	Top-1 Retrieved	Top-2 Retrieved	Top-3 Retrieved
PERSISTENT PATCHY RIGHT BASILAR AIRSPACE OPACITY AND SMALL RIGHT PLEURAL EFFUSION.	PERSISTENT PERIPHERAL RIGHT BASILAR CONSOLIDATION AND SMALL RIGHT PLEURAL EFFUSION.	MODERATE BILATERAL PLEURAL EFFUSIONS, BIBASILAR OPACITIES ATELECTASIS VERSUS CONSOLIDATION.	BILATERAL PLEURAL EFFUSIONS AND BILATERAL BIBASAL OPACITIES THAT COULD REPRESENT CONSOLIDATION.
SMALL BILATERAL PLEURAL EFFUSIONS.	SMALL BILATERAL PLEURAL EFFUSIONS.	SMALL BILATERAL PLEURAL EFFUSIONS.	SMALL BILATERAL PLEURAL EFFUSIONS.
2.NEW SMALL BILATERAL PLEURAL EFFUSIONS.	Small bilateral pleural effusions, left greater than right.	SMALL BILATERAL PLEURAL EFFUSIONS.	Small bilateral pleural effusions, left greater than right.
MODERATE BILATERAL PLEURAL EFFUSIONS ARE AGAIN SEEN.	UNCHANGED MODERATE LEFT AND SMALL RIGHT PLEURAL EFFUSIONS.	MODERATE BILATERAL PLEURAL EFFUSIONS, INCREASED ON THE RIGHT.	A LEFT PLEURAL EFFUSION PERSISTS AND MAY BE SOMEWHAT LARGER.
PERSISTENT MODERATE RIGHT SUBPLEURAL EFFUSION AND SMALL LEFT PLEURAL EFFUSION WITH BIBASILAR AIR SPACE OPACITIES REFLECTING ATELECTASIS OR CONSOLIDATIONS.	1.TWO VIEWS OF THE CHEST DEMONSTRATE BILATERAL PLEURAL EFFUSIONS, RIGHT GREATER LEFT WITH A RIGHT LOWER LOBE OPACITY POSSIBLY REPRESENTING CONSOLIDATION VERSUS ATELECTASIS	NEW PLEURAL PIGTAIL DRAIN AT THE LEFT LUNG BASE WITH STABLE LEFT-SIDED PLEURAL EFFUSION AND LEFT BASILAR OPACITY CONSISTENT WITH ATELECTASIS AND/OR CONSOLIDATION.	STABLE RIGHT PLEURAL EFFUSION WITH ASSOCIATED RIGHT BASILAR ATELECTASIS VERSUS CONSOLIDATION.

Table 16: Qualitative retrieval examples for the CheXpert dataset on the clinical concept pleural effusion. For each query report, we show the top-3 retrieved reports based on cosine similarity of the overall embedding.

Query	Top-1 Retrieved	Top-2 Retrieved	Top-3 Retrieved
2.MILD CARDIOMEGALY.	2.MILD CARDIOMEGALY.	4.MILD CARDIOMEGALY.	6.MILD CARDIOMEGALY.
APPEARANCE OF INCREASED CARDIOMEGALY, THOUGH THIS MAY BE BECAUSE OF SLIGHT LEFTWARD ROTATION.	2.STABLE CARDIOMEDIASTINAL SILHOUETTE WITH CARDIOMEGALY AND MILDLY TORTUOUS DESCENDING AORTA.	THIS IS ASSOCIATED WITH BORDERLINE CARDIOMEGALY.	5.BORDERLINE CARDIOMEGALY, STABLE.
STABLE CARDIOMEDIASTINAL SILHOUETTE WITH REDEMONSTRATION OF CARDIOMEGALY, TORTUOUS AORTA, AND PROMINENT SUPERIOR MEDIASTINUM.	7.PERSISTENT MARKED CARDIOMEGALY.	2.PERSISTENT MARKED CARDIOMEGALY.	2.MARKED CARDIOMEGALY, STABLE .
PERSISTENT MODERATE CARDIOMEGALY.	PERSISTENT MODERATE CARDIOMEGALY.	PERSISTENT MODERATE CARDIOMEGALY.	THERE IS PERSISTENT MODERATE CARDIOMEGALY.
THERE IS PERSISTENT MILD CARDIOMEGALY.	MILD CARDIOMEGALY IS AGAIN NOTED.	MILD CARDIOMEGALY IS AGAIN NOTED.	MILD CARDIOMEGALY IS AGAIN NOTED.

Table 17: Qualitative retrieval examples for the CheXpert dataset on the clinical concept cardiomegaly. For each query report, we show the top-3 retrieved reports based on cosine similarity of the overall embedding.

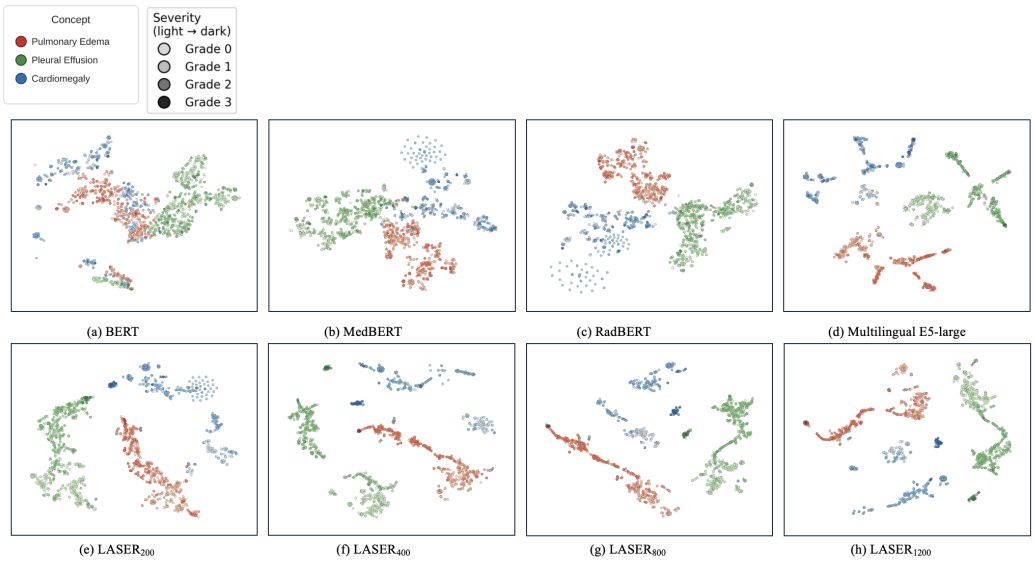


Figure 3: Latent space sentence representations on the CheXpert dataset for the baselines BERT, BioClinical BERT, RadBERT and Multilingual E5-large as well as our model LASER with different numbers of synthetic sentences used for training.

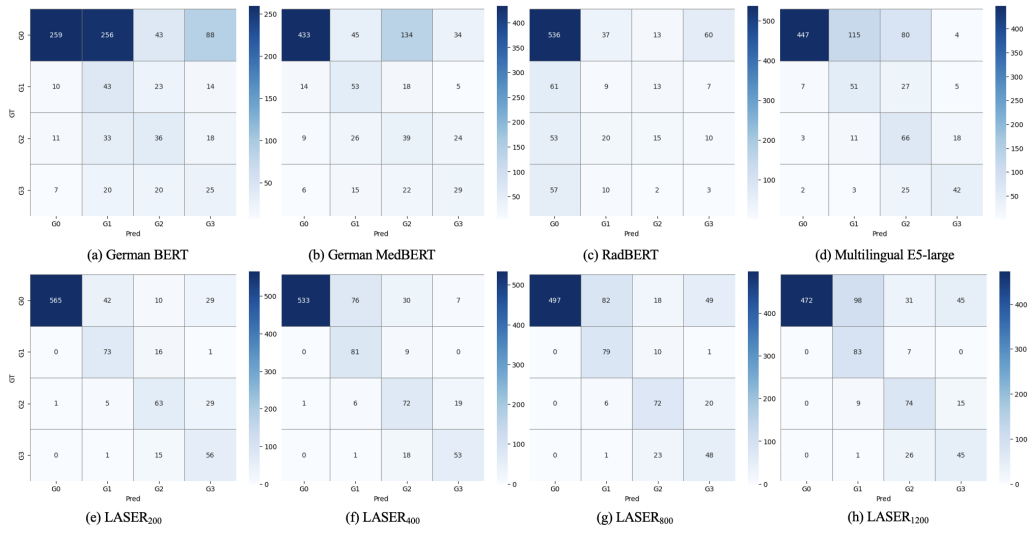


Figure 4: Confusion matrices for the k-nearest neighbor severerity classification on the cardiac dataset. Results are shown for the baselines german BERT, german MedBERT, RadBERT and Multilingual E5-large as well as for our method LASER with different numbers of synthetic sentences used for training.

matrices for severity classification per concept on the CheXpert dataset, complementing the results in Tab.2.

A.4 Reproducibility

Additional details to support the reproducibility of our results, including implementation specifics, training hyperparameters, and the generation process for synthetic training sentences are given below.

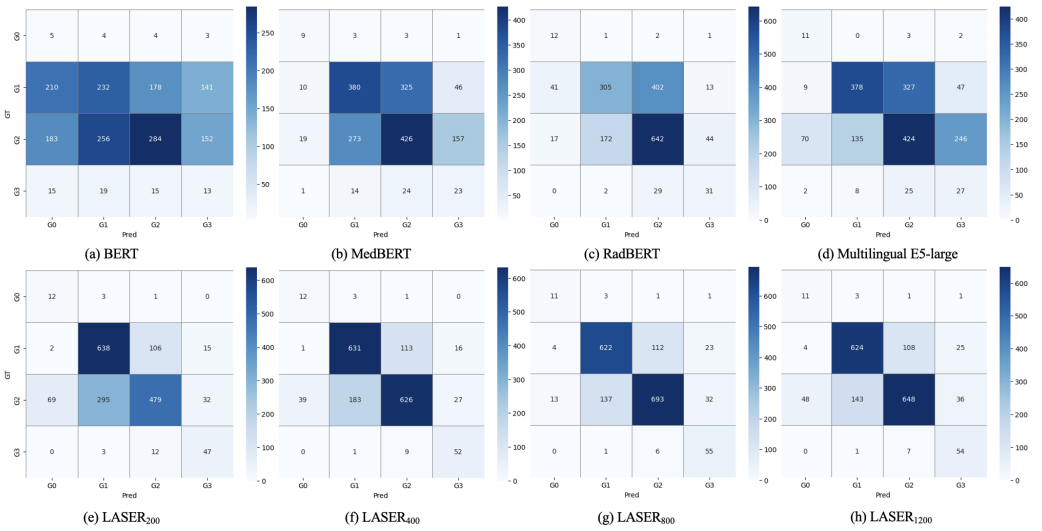


Figure 5: Confusion matrices for the k-nearest neighbor severerity classification on the CheXpert dataset. Results are shown for the baselines BERT, BioClinicalBERT, RadBERT and Multilingual E5-large as well as for our method LASER with different numbers of synthetic sentences used for training.

A.4.1 Implementation Details

We report the hardware setup and training hyperparameters used across all experiments and ablation studies. Unless otherwise specified in the main text, the values listed in Tab.18 apply to all reported configurations. All experiments were conducted on a single NVIDIA A100 GPU (40GB).

Hyperparameter	Value
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$, weight decay $= 0.01$)
Learning rate	1×10^{-6}
LR schedule	ReduceLROnPlateau (patience=5, factor=0.5, mode='min')
Epochs	200
Batch size	32 (internal) / 24 (CheXpert)
Loss weights	$\lambda_{MNRL} = \lambda_{HN} = \lambda_{Tri} = \lambda_{Cos} = 1$
k (kNN)	10
Seeds	42, 43, 44, 45, 46

Table 18: Training hyperparameters.

A.4.2 Synthetic Sentence Generation

For reproducibility purposes, the prompts used to generate the synthetic sentences are given below. Tab.19 show the prompts to generate german sentences for the internal cardiac dataset and Tab.20 those to generate the english sentences for the CheXpert dataset.

Concept	Prompt
Calcification	Generiere synthetische Sätze, die Sätze aus kardiologischen CT befunden (Radiologieberichten) zum klinischen Konzept Verkalkung der Koronararterien mit vier verschiedenen Schweregraden (0-3 für negativ, gering, mäßig, ausgeprägt). Verwende unterschiedliche Formulierungen und unterschiedliche Lokalisationen (RCA, LAD, CX). Erzeuge auch Sätze, bei denen mehrere Arterien unterschiedliche Schweregrade aufweisen und der Gesamtschweregrad dem höheren entspricht. Generiere sie im folgenden YAML-Dateiformat: entries: ”sentence”: severity ”sentence”: severity (also Sätze als strings mit ””). Generiere 4 YAML-Dateien, eine mit 50 Sätzen pro Schweregrad, 100 Sätzen pro Schweregrad, 200 Sätzen pro Schweregrad und 300 pro Schweregrad (diese können verschachtelte Teilmengen voneinander sein). Generiere außerdem eine Validierungs-YAML-Datei mit 40 Sätzen pro Schweregrad (unterschiedlich von allen Sätzen in den anderen Sets).
Stenosis	Generiere synthetische Sätze, die Sätze aus kardiologischen CT befunden (Radiologieberichten) zum klinischen Konzept Stenose der Koronararterien mit vier verschiedenen Schweregraden (0-3 für negativ, gering, mäßig, ausgeprägt). Verwende unterschiedliche Formulierungen und unterschiedliche Lokalisationen (RCA, LAD, CX). Erzeuge auch Sätze, bei denen mehrere Arterien unterschiedliche Schweregrade aufweisen und der Gesamtschweregrad dem höheren entspricht. Generiere sie im folgenden YAML-Dateiformat: entries: ”sentence”: severity ”sentence”: severity (also Sätze als strings mit ””). Generiere 4 YAML-Dateien, eine mit 50 Sätzen pro Schweregrad, 100 Sätzen pro Schweregrad, 200 Sätzen pro Schweregrad und 300 pro Schweregrad (diese können verschachtelte Teilmengen voneinander sein). Generiere außerdem eine Validierungs-YAML-Datei mit 40 Sätzen pro Schweregrad (unterschiedlich von allen Sätzen in den anderen Sets).
Aorta/TP	Generiere synthetische Sätze, die Sätze aus kardiologischen CT befunden (Radiologieberichten) zum klinischen Konzept erweiterte Aorta oder Truncus pulmonalis mit vier unterschiedlichen Schweregraden (0–3 für keine Dilatation, geringe Dilatation, mäßige Dilatation, ausgeprägte Dilatation). Erzeuge außerdem Sätze, bei denen nur Teile der Aorta oder nur der Truncus pulmonalis dilatiert sind. Generiere sie im folgenden YAML-Dateiformat: entries: ”sentence”: severity ”sentence”: severity (also Sätze als strings mit ””). Generiere 4 YAML-Dateien, eine mit 50 Sätzen pro Schweregrad, 100 Sätzen pro Schweregrad, 200 Sätzen pro Schweregrad und 300 pro Schweregrad (diese können verschachtelte Teilmengen voneinander sein). Generiere außerdem eine Validierungs-YAML-Datei mit 40 Sätzen pro Schweregrad (unterschiedlich von allen Sätzen in den anderen Sets).
Myocardium	Generiere synthetische Sätze, die Sätze aus kardiologischen CT befunden (Radiologieberichten) zum klinischen Konzept auffälliges Myokard mit vier unterschiedlichen Schweregraden (0–3 für keine Auffälligkeit, hypertrophes/verdicktes Myokard, mäßige Myokardpathologie, ausgeprägte Auffälligkeit, d. h. Infarkt, ausgeprägter Funktionsverlust oder Narbe). Generiere sie im folgenden YAML-Dateiformat: entries: ”sentence”: severity ”sentence”: severity (also Sätze als strings mit ””). Generiere 4 YAML-Dateien, eine mit 50 Sätzen pro Schweregrad, 100 Sätzen pro Schweregrad, 200 Sätzen pro Schweregrad und 300 pro Schweregrad (diese können verschachtelte Teilmengen voneinander sein). Generiere außerdem eine Validierungs-YAML-Datei mit 40 Sätzen pro Schweregrad (unterschiedlich von allen Sätzen in den anderen Sets).

Table 19: Prompts used to generate the german synthetic sentences for fine-tuning the text encoder for the internal cardiac dataset.

Concept	Prompt
Pulmonary Edema	Generate synthetic sentences representing sentences from chest X-ray radiology reports concerning the clinical concept pulmonary edema with four different severities (0-3 for no edema, mild edama, moderate edema, severe edema). Use different wordings and different locations (left or right lung). Generate also sentences, where both sides have different severity levels and the overall severity is the larger one. Generate them in a the following yaml file format: entries: ”sentence”: severity ”sentence”: severity. Generate 4 yaml files, one with 50 sentences per severity, 100 sentences per severity, 200 sentences per severity and 300 per severity (this can be nested subsets from each other). Also generate a validation yaml file with 40 sentences per severity (different from all sentences in the other sets).
Pleural Effusion	Generate synthetic sentences representing sentences from chest X-ray radiology reports concerning the clinical concept pleural effusion with four different severities (0-3 for negative, small, moderate, large). Use different wordings and different locations (left or right lung). Generate also sentences, where both sides have different severity levels and the overall severity is the larger one. Generate them in a the following yaml file format: entries: ”sentence”: severity ”sentence”: severity. Generate 4 yaml files, one with 50 sentences per severity, 100 sentences per severity, 200 sentences per severity and 300 per severity (this can be nested subsets from each other). Also generate a validation yaml file with 40 sentences per severity (different from all sentences in the other sets).
Cardiomegaly	Generate synthetic sentences representing sentences from chest X-ray radiology reports concerning the clinical concept Cardiomegaly with four different severities (0-3 for normal heart size, mildly enlarged heart, moderately enlarged heart, severely enlarged heart). Use different wordings. Generate them in a the following yaml file format: entries: ”sentence”: severity ”sentence”: severity. Generate 4 yaml files, one with 50 sentences per severity, 100 sentences per severity, 200 sentences per severity and 300 per severity (this can be nested subsets from each other). Also generate a validation yaml file with 40 sentences per severity (different from all sentences in the other sets).

Table 20: Prompts used to generate the english synthetic sentences for fine-tuning the text encoder for the CheXpert dataset.