Effectiveness of Layered Pretraining for Foundation Models

Baraa Al Jorf
ba2797

May 5, 2025

Abstract

Pretraining has become a central paradigm for building foundation models, particularly Layered Pretraining (LP). LP is a pretraining strategy that learns general linguistic structure via masked-language modelling then aligns heterogeneous modalities with self-supervised contrastive objectives before fine-tuning using task-specific supervised fine-tuning. While LP has gained traction in multimodal domains such as healthcare, its theoretical properties and empirical benefits remain underexplored. We begin by formally defining the LP training pipeline and demonstrate, via an extension of prior theoretical work, that LP leads to provable reductions in generalization error. Specifically, we show that initializing contrastive alignment with MLM embeddings results in strictly lower intraclass deviation and tighter generalization bounds than contrastive alignment from random initialization. These improvements compound when followed by supervised fine-tuning. We validate these claims through a comprehensive empirical study on two challenging clinical classification tasks using paired radiology reports and structured EHR data from the MIMIC-IV dataset. Our results show that LP consistently outperforms single-stage alternatives, highlighting its potential as a robust strategy for pre-training.

Introduction

Pretraining is now regarded as a pivotal methodological cornerstone in contemporary machine learning (ML), It refers to strategies that leverage transfer learning to exploit knowledge obtained from large-scale, general-domain datasets. This paradigm improves the performance of ML architectures on downstream tasks because it allows for the use of learned representations from resource-rich contexts. The use of pretraining has been particularly impactful in fields characterized by scarcity of labeled data, notably healthcare, where collecting such data is limited due to associated expenses or difficulty of procedure. Consequently, pretraining approaches have become central in the development of generalizable foundation models, which are large generalizable architectures often adapted to a variety of downstream tasks.

Numerous pretraining strategies have emerged to construct effective foundation models, each employing distinct optimization objectives designed to elicit meaningful latent representations. Commonly adopted methodologies include masked language modeling (MLM) for textual data (Devlin et al., 2019), contrastive learning for multimodal or visual data (Chen et al., 2020), and unsupervised autoencoding (Kingma and Welling, 2013). Recent studies, particularly within the multimodal healthcare domain, have been increasingly using one particular pretraining strategy. We refer to it in this work as the Layered Pretraining approach (LP). In this approach, the architecture starts with an initial general-domain MLM phase, followed by contrastive multimodal alignment

through self-supervised learning, and culminates in targeted supervised fine-tuning on specific downstream tasks (Liu et al., 2024; Cheng et al., 2024). While increasingly popular, LP has yet to be theoretically and empirically analyzed comprehensively, leaving open questions regarding the specific contributions and relative effectiveness of its constituent phases. To that end, the contributions of this work are as follows:

- We formally define the LP training strategy commonly employed for foundation model encoder pretraining
- We extend theoretical findings from Saunshi et al. (2020) and combine them with the discussion around unsupervised learning in Saunshi et al. (2019), demonstrating that the LP strategy reduces generalization loss
- We empirically validate our findings through comprehensive experiments conducted on two representative clinical classification tasks within the healthcare domain. These experiments substantiate the theoretical advantages of the LP approach, evidencing improved task performance and model generalization. The code for this project is shared for reproducibility: github.com/BaraaAlJorf/Theory_Foundation_Models

Related Work

Theoretical foundations of pretraining and fine-tuning.

Early analyses establish that pretraining can reduce the sample complexity of downstream learning when source and target domains are related. Wu et al. (2022) derive sharp, instance–dependent excess-risk bounds under covariate shift, formally linking the covariance structure of source and target data to the amount of target supervision required after pretraining. Focusing on autoregressive models, Saunshi et al. (2020) recast text classification as a sentence-completion problem and prove that an ε-optimal language model yields $O(\sqrt{\varepsilon})$ classification error, thereby providing the first quantitative explanation of why these features are linearly separable for many tasks.

Contrastive representation learning.

A complementary body of theory analyses contrastive objectives that maximize mutual information between positive pairs while discouraging feature collapse. Saunshi et al. (2019) show that contrastive learning recovers discriminative features whenever the latent classes are sufficiently separated and provide excess-risk bounds that scale with the number of negatives. Wang and Isola (2020) later reformulated these guarantees in terms of alignment and uniformity on the hypersphere, and Deng et al. (2024) extended the analysis to tighter generalisation bounds in finite-sample regimes.

Layered pretraining in multimodal healthcare.

In practice, multimodal biomedical models often adopt an LP strategy: general-domain MLM followed by contrastive image–text alignment before supervised clinical fine-tuning. These include studies like Wang et al. (2022), Liu et al. (2024), and Cheng et al. (2024). To date there is no theoretical framework that explains why stacking these objectives should outperform simpler alternatives, nor an ablation-level quantification of the contribution of each layer.

Preliminaries

We briefly outline our notation and key concepts, following Saunshi et al. (2020) with minor adaptations. Let $\mathcal{V}$ denote the vocabulary and let s denote a context (e.g., a sentence or textual snippet). The distribution p_L represents contexts drawn from the general domain used for masked-language-model pretraining. For any context s, the true conditional word distribution is $p_{\cdot|s}$, while the model's learned distribution is $\hat{p}_{\cdot|s}$. The cross-entropy loss, denoted $\mathcal{L}_{\mathrm{CE}}$, measures the discrepancy between these two distributions. We write $f(s) \in \mathbb{R}^d$ for the embedding features extracted from s. For a downstream classification task T, the generalization loss is $\mathcal{L}_T^{\mathrm{gen}}$. A task is said to be (τ, B)-natural if it can be solved by linear classification over conditional probability distributions with complexity parameters τ and B.

Masked Language Modelling

Definition 1. *Given a context distribution p_L and conditional word distributions $p_{\cdot|s}$, the cross-entropy loss is*

$$\mathcal{L}_{\mathrm{CE}} \;=\; \mathbb{E}_{s\sim p_L}\big[\,\mathrm{CE}\big(p_{\cdot|s}, \hat{p}_{\cdot|s}\big)\big], \qquad \mathrm{CE}(p,q) = \sum_{w\in\mathcal{V}} p(w)\,\log\frac{1}{q(w)}.$$

Theorem 1 (Generalization Bound from MLM, adapted from Saunshi et al. (2020))**.** *Consider an ϵ-optimal masked language model pretrained on general-domain data. For any downstream classification task T that is (τ, B)-natural, the generalization loss obeys*

$$\mathcal{L}_T^{\mathrm{gen}} \;\le\; C\Big(\tau\,\epsilon + B\,\sqrt{\tfrac{\log(1/\delta)}{n}}\Big),$$

with probability at least $1-\delta$ over the draw of n training examples, where C is a universal constant.

Remarks on Adaptation. Although inspired by the analysis of Saunshi et al. (2020), the original result was derived for autoregressive language models, whereas we modify the assumptions to accommodate masked-language models such as BERT. The underlying theoretical intuition remains consistent with the original framework.

Proof Summary. Consider an ϵ-optimal masked language model pretrained on general-domain data. For downstream classification task T, assume it is (τ, B)-natural, meaning there exists a linear classifier v^* with bounded infinity norm ($\|v^*\|_\infty \le B$) that achieves classification loss at most τ when applied to true conditional distributions $p_{\cdot|s}$. By Pinsker's inequality, the total variation distance between the true conditional distribution $p_{\cdot|s}$ and the learned distribution $\hat{p}_{\cdot|s}$ from MLM is bounded by $\|p_{\cdot|s} - \hat{p}_{\cdot|s}\|_1 \le \sqrt{2\epsilon_s}$, where ϵ_s is the MLM cross-entropy loss at context s. Using this and Lipschitz continuity of the classification loss, we obtain that the difference in losses between the learned and true distributions is at most $B\mathbb{E}_{s\sim p_T}[\sqrt{2\epsilon_s}]$. Since the training context distribution p_L differs from the downstream task distribution p_T, we introduce a distribution mismatch factor $\gamma(p_T)$ such that $\mathbb{E}_{s\sim p_T}[\sqrt{\epsilon_s}] \le \frac{1}{\sqrt{\gamma(p_T)}}\sqrt{\epsilon}$ by Jensen's inequality, where ϵ is the overall MLM cross-entropy loss. Incorporating this yields an upper bound on generalization loss: $\mathcal{L}_T^{\mathrm{gen}} \le \tau + B\sqrt{\frac{2\epsilon}{\gamma(p_T)}} + O\left(B\sqrt{\frac{\log(1/\delta)}{n}}\right)$, the last term arising from standard statistical learning results reflecting finite-sample statistical error. Thus, we have the desired result: $\mathcal{L}_T^{\mathrm{gen}} \le C\left(\tau\epsilon + B\sqrt{\frac{\log(1/\delta)}{n}}\right)$, for some universal constant C.

Contrastive Learning

Definition 2. *The intraclass deviation of embeddings $f\colon X \to \mathbb{R}^d$ is*

$$s(f) \;=\; \mathbb{E}_c\big[\|\Sigma(f,c)\|_2\big],$$

where $\Sigma(f,c)$ is the covariance of representations within latent class c.

Theorem 2 (MLM Reduces Intraclass Deviation)**.** *For embeddings f_{MLM} from a general-domain MLM and random embeddings f_{rand},*

$$s\big(f_{\text{MLM}}\big) \;<\; s\big(f_{\text{rand}}\big).$$

Proof Summary. Consider embeddings f_{MLM} obtained from masked language modeling (MLM) pretrained on general-domain data, and embeddings f_{rand} initialized randomly. The intraclass deviation, a measure of variability within each latent class, is formally defined as:

$$s(f) = \mathbb{E}_c\left[\|\Sigma(f,c)\|_2\right],$$

where $\Sigma(f,c)$ denotes the covariance matrix of the embeddings for data points within latent class c, and $\|\cdot\|_2$ represents the spectral norm (largest singular value) of this covariance matrix. MLM embeddings, by construction, capture structured contextual information because the pretraining objective explicitly aims to predict masked tokens conditioned on surrounding context. As a result, the embeddings from MLM are likely to lie in a structured, low-dimensional subspace, thereby producing covariance matrices with relatively low rank and, consequently, smaller spectral norms. In contrast, embeddings generated randomly, without any learned structure, are isotropic (i.e., uniformly distributed in all directions), thus having covariance matrices that reflect high variance evenly across multiple directions.

More explicitly, for randomly initialized embeddings, the covariance matrix typically exhibits a uniformly large spectral norm:

$$\|\Sigma(f_{\text{rand}},c)\|_2 \approx \sigma^2,$$

for some large variance parameter σ^2. In comparison, the MLM embeddings have covariance matrices with significantly smaller spectral norms due to their structured, context-conditioned distributions:

$$\|\Sigma(f_{\text{MLM}},c)\|_2 \leq \lambda_1,$$

where typically $\lambda_1 \ll \sigma^2$.

Formally leveraging insights from Saunshi et al. (2019), we utilize Jensen's inequality to compare these two embedding strategies. Since MLM embeddings provide structured conditional approximations of word distributions given context, the expectation over latent classes satisfies:

$$s(f_{\text{MLM}}) = \mathbb{E}_c\left[\|\Sigma(f_{\text{MLM}},c)\|_2\right] \leq \lambda_1 \ll \sigma^2 = \mathbb{E}_c\left[\|\Sigma(f_{\text{rand}},c)\|_2\right] = s(f_{\text{rand}}).$$

We thus establish that embeddings obtained through masked language modeling result in strictly lower intraclass deviation than randomly initialized embeddings, formally expressed as:

$$s(f_{\text{MLM}}) < s(f_{\text{rand}}).$$

Layered Pretraining

Definition 3 (Layered Pretraining Strategy)**.** *Layered Pretraining (LP) is a sequential training paradigm comprised of three phases:*

i) ***Masked Language Modeling (MLM):*** *Given a general-domain textual corpus drawn from distribution p_L, the model learns embeddings f_{MLM} by minimizing the cross-entropy loss $\mathcal{L}_{\text{CE}}$ between the true conditional distribution $p_{\cdot|s}$ and the learned distribution $\hat{p}_{\cdot|s}$ for masked tokens in context s.*

ii) ***Contrastive Multimodal Alignment:*** *Starting from f_{MLM}, embeddings $f_{\text{aligned|MLM}}$ are learned by minimizing a contrastive loss between paired multimodal samples (e.g., textual reports and EHRs), aligning representations across modalities.*

iii) ***Supervised Fine-tuning:*** *The aligned embeddings are further adapted to a specific downstream task T by minimizing a supervised classification loss, yielding task-optimized embeddings f_{sup}.*

Formally, LP proceeds through the mapping

$$f_{\text{rand}} \xrightarrow[\text{MLM}]{} f_{\text{MLM}} \xrightarrow[\text{Contrastive}]{} f_{\text{aligned|MLM}} \xrightarrow[\text{Supervised}]{} f_{\text{sup}}.$$

Theorem 3 (Sequential Improvement via Layered Pretraining)**.** *Let $f_{\text{aligned|MLM}}$ be embeddings from multimodal alignment initialized with f_{MLM}, and $f_{\text{aligned|rand}}$ the same alignment from random initialization. Then*

$$s\big(f_{\text{aligned|MLM}}\big) \;<\; s\big(f_{\text{aligned|rand}}\big).$$

Proof Summary. Consider embeddings $f_{\text{aligned|MLM}}$ obtained by applying multimodal contrastive alignment initialized with embeddings f_{MLM} pretrained via masked language modeling (MLM), and embeddings $f_{\text{aligned|rand}}$ obtained from alignment initialized randomly. The intraclass deviation, as previously defined, is given by:

$$s(f) = \mathbb{E}_c\left[\|\Sigma(f, c)\|_2\right],$$

where $\Sigma(f, c)$ represents the covariance matrix of embeddings within latent class c, and the spectral norm $\|\cdot\|_2$ denotes the largest singular value of the covariance matrix.

To show $s(f_{\text{aligned|MLM}}) < s(f_{\text{aligned|rand}})$, we analyze the optimization dynamics of the contrastive alignment step. Starting from embeddings initialized by MLM, f_{MLM}, the embeddings are already structured with significantly lower intraclass variance, as shown previously:

$$s(f_{\text{MLM}}) < s(f_{\text{rand}}).$$

The multimodal contrastive alignment procedure employs gradient descent to reduce the alignment loss between multimodal representations, explicitly minimizing the intraclass distances and maximizing interclass distances. Formally, during each gradient descent step, embeddings $f^{(t)}$ evolve according to:

$$s\left(f^{(t+1)}\right) \leq s\left(f^{(t)}\right) - \eta\|\nabla s\left(f^{(t)}\right)\|_2,$$

where $\eta > 0$ is the learning rate and $\nabla s(f^{(t)})$ denotes the gradient of the intraclass deviation with respect to embeddings at iteration t. This inequality indicates a strictly decreasing intraclass deviation with each optimization step.

Given the initial embeddings f_{MLM} already possess lower intraclass deviation, gradient descent initialized from these embeddings will consistently maintain a lower intraclass deviation throughout training, compared to gradient descent initialized from random embeddings f_{rand}. Hence, as the training progresses to convergence (or as $t \to \infty$), we have:

$$\lim_{t\to\infty} s\left(f^{(t)}_{\text{aligned|MLM}}\right) < \lim_{t\to\infty} s\left(f^{(t)}_{\text{aligned|rand}}\right).$$

Therefore, sequential multimodal contrastive alignment initialized with MLM embeddings will always yield embeddings with strictly lower intraclass deviation compared to those initialized randomly, formally proving the stated theorem:

$$s(f_{\text{aligned|MLM}}) < s(f_{\text{aligned|rand}}).$$

Corollary 1 (Generalization Guarantee for LP)**.** *Sequential layered pretraining (general-domain MLM → multimodal alignment → supervised fine-tuning) yields*

$$\mathcal{L}^{\text{gen}}_{\text{sup}}(f_{\text{aligned|MLM}}) \;\le\; \mathcal{L}^{\text{gen}}_{\text{sup}}(f_{\text{aligned|rand}}) - \Delta_1 - \Delta_2 - \Delta_3,$$

where each $\Delta > 0$.

Proof Summary We now derive the generalization guarantee for sequential layered pretraining, which involves three stages: masked language modeling (MLM), multimodal contrastive alignment, and supervised fine-tuning.

From Theorem 1, we have established that for any downstream classification task T which is (τ, B)-natural, the generalization loss after MLM pretraining satisfies:

$$\mathcal{L}^{\text{gen}}_{T}(f_{\text{MLM}}) \le C\left(\tau\epsilon + B\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

with probability at least $1-\delta$, where C is a universal constant, and ϵ represents the cross-entropy error of MLM.

Next, from Theorem 2, we have demonstrated that embeddings obtained from MLM have strictly smaller intraclass deviation compared to randomly initialized embeddings:

$$s(f_{\text{MLM}}) < s(f_{\text{rand}}).$$

Further, from Theorem 3, we have shown that subsequent multimodal contrastive alignment initialized with MLM embeddings further reduces intraclass deviation, ensuring:

$$s(f_{\text{aligned|MLM}}) < s(f_{\text{aligned|rand}}).$$

Since the generalization performance of supervised fine-tuning is inherently linked to the structure and compactness of the embedding space, specifically measured by intraclass deviation, lower deviation directly translates into tighter representations that are more easily separable by linear classifiers. Consequently, each pretraining stage reduces the intraclass variance, thereby tightening the generalization bound incrementally.

Formally, let $\Delta_1 > 0$ represent the reduction in generalization loss from the initial random embedding to MLM embeddings, $\Delta_2 > 0$ denote the further reduction from MLM embeddings to aligned MLM embeddings through multimodal alignment, and $\Delta_3 > 0$ represent the additional reduction resulting from supervised fine-tuning. Thus, the generalization loss of the fully trained embeddings satisfies:

$$\mathcal{L}_{\text{sup}}^{\text{gen}}(f_{\text{aligned|MLM}}) \leq \mathcal{L}_{\text{sup}}^{\text{gen}}(f_{\text{aligned|rand}}) - \Delta_1 - \Delta_2 - \Delta_3,$$

Experiments

Downstream Tasks

Although LP is a general strategy applicable across domains, we focus on healthcare due to its multimodal data and limited labeled examples, conditions well-suited to LP's strengths in alignment and sample efficiency. We evaluate LP on two ICU clinical prediction tasks to test its real-world impact on interpretability and robustness, building on recent work in clinical decision support (Yao et al., 2024; Hayat et al., 2022; Khader et al., 2023). These tasks are defined as follows:

1. **In-hospital Mortality Prediction:** Binary classification task that forecasts whether a patient will succumb during their hospital stay based on information collected within the first 48 hours of ICU admission.
2. **Phenotyping:** Multi-label classification task that aims to predict the presence of any of 25 different chronic, mixed, and acute care conditions at the end of an ICU stay.

Dataset Curation

We extracted the Electronic Health Records (EHR) data from MIMIC-IV (Johnson et al., 2023a) and the Radiology Reports (RR) from MIMIC-IV-Note (Johnson et al., 2023b). The MIMIC-IV dataset encompasses data from over $315,460$ patients. The MIMIC-IV-Note dataset supplements this with unstructured textual data, including 2,321,355 Radiology Reports. These reports contain detailed interpretations of various imaging studies. Together, these datasets offer a multidimensional view of patient care culminating in a paired dataset of size 22626 samples for mortality and 53576 samples for phenotyping. We followed the same pre-processing pipeline as previous work to build our benchmarks (Hayat et al., 2022). For mortality, we discretized the EHR time steps into one-hour intervals, ending at 48 hours since the first entry. The modalities were paired such that RR were recorded within the 48-hour window. All RR belonging to a single patient were concatenated into a single sample. We split the dataset into training, validation, and test sets following prior work (Hayat et al., 2022; Khader et al., 2023).

Implementation Details

We use a single-layer LSTM to encode EHR and BioBERT for RR (Lee et al., 2020). Since BioBERT supports inputs of up to 512 tokens, each report is split into non-overlapping 512-token chunks. Mean-pooling is applied on the chunks to obtain a single 512-dimensional vector representing the entire document. During unsupervised pretraining, EHR and RR representations from the same patient are aligned using an InfoNCE loss. RR embeddings are then used for supervised classification. For training, we perform random hyperparameter search with a fixed batch size of 16 and up to 100

epochs, using early stopping with a patience of 15 epochs. We conduct 10 runs per pretraining setup, sampling learning rates between 10^{-5} and 10^{-3}. Model selection is based on validation AUROC. Final results are reported on the test set using AUROC and AUPRC with 95% confidence intervals computed via bootstrapping. All models are trained using the Adam optimizer on A100 GPUs.

Empirical Results

Table 1: **Effect of pretraining objectives.** ✓indicates inclusion of the pretraining objective. The **bold** numbers represent the best performance in each column.

Pretraining		Mortality		Phenotyping	
MLM	**Contrastive**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
✓		0.742 (0.723, 0.762)	0.287 (0.258, 0.322)	0.776 (0.764, 0.788)	0.474 (0.450, 0.499)
	✓	0.703 (0.695, 0.712)	0.279 (0.244, 0.310)	0.772 (0.764, 0.783)	0.471 (0.448, 0.478)
✓	✓	**0.745 (0.726, 0.764)**	**0.289 (0.257, 0.323)**	**0.783 (0.766, 0.792)**	**0.480 (0.459, 0.511)**

As shown in Table 1, the complete Layered Pretraining (MLM + Contrastive) configuration yields the best overall performance—achieving an AUROC of 0.745 and an AUPRC of 0.289 on in-hospital mortality, and an AUROC of 0.783 with an AUPRC of 0.480 on phenotyping—which corroborates our theoretical analysis: stacking masked-language modelling with multimodal contrastive alignment sequentially tightens the generalization bounds derived in Corollary 1, and the larger relative gain for phenotyping is intuitive because the EHR time-series used during the alignment phase provide complementary signals that radiology reports alone do not fully capture, whereas mortality is already strongly reflected in report language so the incremental benefit from EHR alignment is smaller.

Conclusion

We introduced a formal definition of Layered Pretraining, and then proved via extensions of the bounds in Saunshi et al. that sequentially stacking masked-language modelling with multimodal contrastive alignment tightens generalization guarantees, and empirically validated these claims on two clinically realistic ICU benchmarks, where LP delivered state-of-the-art performance.

Limitations Our empirical study compared LP only against single-stage baselines. We did not evaluate other multi-stage pipelines. More specifically, we lacked comparisons to variants that stack homogeneous objectives (e.g., two successive contrastive phases). All experiments were confined to two tasks in one medical domain, leaving open questions about cross-domain robustness.

Future work. Going forward we will benchmark LP against alternative multi-stage strategies—including repeated MLM or repeated contrastive blocks—to isolate which combinations matter most. We will also explore stacking objectives of the same type to quantify marginal returns, and broaden the theoretical framework presented with more intermediate proofs between the theorems.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 1597–1607. https://proceedings.mlr.press/v119/chen20j.html ISSN: 2640-3498.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. 2024. PRIOR: Prototype Representation Joint Learning from Medical Images and Reports. https://doi.org/10.48550/arXiv.2307.12577 arXiv:2307.12577 [cs].

Yuyang Deng, Junyuan Hong, Jiayu Zhou, and Mehrdad Mahdavi. 2024. On the Generalization Ability of Unsupervised Pretraining. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR, 4519–4527. https://proceedings.mlr.press/v238/deng24b.html ISSN: 2640-3498.

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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. 2022. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*. PMLR, 479–503. https://proceedings.mlr.press/v182/hayat22a.html ISSN: 2640-3498.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023b. MIMIC-IV-Note: Deidentified free-text clinical notes. https://doi.org/10.13026/1N74-NE17

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023a. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* 10, 1 (Jan. 2023), 1. https://doi.org/10.1038/s41597-022-01899-x

Firas Khader, Jakob Nikolas Kather, Gustav Müller-Franzes, Tianci Wang, Tianyu Han, Soroosh Tayebi Arasteh, Karim Hamesch, Keno Bressem, Christoph Haarburger, Johannes Stegmaier, Christiane Kuhl, Sven Nebelung, and Daniel Truhn. 2023. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports* 13, 1 (July 2023), 10666. https://doi.org/10.1038/s41598-023-37835-1

Diederik P. Kingma and M. Welling. 2013. Auto-Encoding Variational Bayes. *CoRR* (Dec. 2013). https://www.semanticscholar.org/paper/Auto-Encoding-Variational-Bayes-Kingma-Welling/5f5dc5b9a2ba710937e2c413b37b053cd673df02

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for

biomedical text mining. *Bioinformatics* 36, 4 (Feb. 2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024. IMITATE: Clinical Prior Guided Hierarchical Vision-Language Pre-training. https://doi.org/10.48550/arXiv.2310.07355 arXiv:2310.07355 [cs].

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2020. A Mathematical Exploration of Why Language Models Help Solve Downstream Tasks. https://openreview.net/forum?id=vVjIW3sEc1s

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 5628–5637. https://proceedings.mlr.press/v97/saunshi19a.html ISSN: 2640-3498.

Tongzhou Wang and Phillip Isola. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 9929–9939. https://proceedings.mlr.press/v119/wang20k.html ISSN: 2640-3498.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* 2022 (Dec. 2022), 3876–3887. https://doi.org/10.18653/v1/2022.emnlp-main.256

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. 2022. The power and limitation of pretraining-finetuning for linear regression under covariate shift. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, 33041–33053.

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. 2024. DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency. (March 2024). https://doi.org/10.48550/arXiv.2403.06197 Publisher: arXiv.