BiCAL: Bi-directional Contrastive Active Learning for Clinical Report Generation

Tianyi Wu[1], **Jingqing Zhang**[1, 2], **Wenjia Bai**[1], **Kai Sun**[1]
[1]Imperial College London [2]Pangaea Data
[1]{andrew.wu22, jingqing.zhang15, w.bai, k.sun}@imperial.ac.uk
[2]jzhang@pangaeadata.ai

Abstract

State-of-the-art performance by large pre-trained models in computer vision (CV) and natural language processing (NLP) suggests their potential for domain-specific tasks. However, training these models requires vast amounts of labelled data, a challenge in many domains due to the cost and expertise required for data labelling. Active Learning (AL) can mitigate this by selecting minimal yet informative data for model training. While AL has been mainly applied to single-modal tasks in the fields of NLP and CV, its application in multi-modal tasks remains underexplored. In this work, we proposed a novel AL strategy, **Bi**directional **C**ontrastive **A**ctive **L**earning strategy (BiCAL), that used both image and text latent spaces to identify contrastive samples to select batches to query for labels. BiCAL was robust to class imbalance data problems by its design, which is a problem that is commonly seen in training domain-specific models. We assessed BiCAL's performance in domain-specific learning on the clinical report generation tasks from chest X-ray images. Our experiments showed that BiCAL outperforms State-of-the-art methods in clinical efficacy metrics, improving recall by 2.4% and F1 score by 9.5%, showcasing its effectiveness in actively training domain-specific multi-modal models.

1 Introduction

Active Learning (AL) is a branch of machine learning that aims to select a small set of the most informative data to annotate for model training (Settles, 2009). This technique allows the model to achieve optimal performance while lowering the cost of annotation. Moreover, by actively selecting data to train on, a model trained under active learning can sometimes surpass the performance that is trained on the full dataset. AL has shown its great potential in the field of natural language processing (NLP) (Shelmanov et al., 2021; Dor et al., 2020; Shen et al., 2017; Margatina et al., 2021a) and computer vision (CV) (Slade and Branson, 2022; Takezoe et al., 2023). However, relatively few have explored the application of active learning in a multi-modal setting.

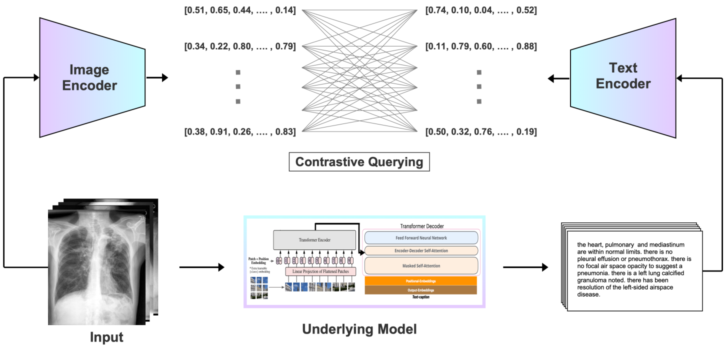
[0.51, 0.65, 0.44, ... , 0.14]
[0.34, 0.22, 0.80, ... , 0.79]
[0.38, 0.91, 0.26, ... , 0.83]
[0.74, 0.10, 0.04, ... , 0.52]
[0.11, 0.79, 0.60, ... , 0.88]
[0.50, 0.32, 0.76, ... , 0.19]
Image Encoder
Text Encoder
Contrastive Querying
Input
Transformer Encoder
Patch + Position Embedding
Linear Projection of Flattened Patches
Transformer Decoder
Feed Forward Neural Network
Encoder-Decoder Self-Attention
Masked Self-Attention
Positional-Embeddings
Output-Embeddings
Text-caption
Underlying Model
the heart, pulmonary and mediastinum are within normal limits. there is no pleural effusion or pneumothorax. there is no focal air space opacity to suggest a pneumonia. there is a left lung calcified granuloma noted. there has been resolution of the left-sided airspace disease.

Figure 1: Flowchart of the querying process of BiCAL: Image is passed to imaged encoder to obtain image embeddings, and the underlying training model to generate reports. Reports generated are passed to a text encoder to generate text embeddings. Together two embeddings are compared and the contrastiveness of each data point is calculated and queried. Refer detail to Algorithm 1.

In addition, as the capabilities of general large-pretrained models arise (Bai et al., 2023; OpenAI, 2023), a rising interest has been seen in fine-tuning them to become domain-specific models. However, when training models in specific domains, obtaining quality labelled data is challenging due to the domain expertise required for accurate annotation, which is costly in both time and money. This motivates us to explore active learning's application in the domain-specific setting. We identify that in domain-specific active learning, there exists one key challenge – class imbalance in datasets is often seen in domain-specific settings, existing AL methods struggle to actively select samples that have less population but may be more important – in medicine, common (healthy) samples often out populate rare (unhealthy) samples. Models trained

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 328–341
August 16, 2024. ©2024 Association for Computational Linguistics

under such active learning strategies converge on the commonly seen samples and perform poorly in identifying rare sickness cases.

In this study, we introduce a novel AL strategy **Bi**directional **C**ontrastive **A**ctive **L**earning strategy (BiCAL) that is tailored to address the challenge in domain-specific active learning. We assess BiCAL and other established AL methods on clinical report generation from chest X-ray images. Our key contributions are:

1. We propose a novel AL strategy BiCAL that is able to select rare but important cases inherently to be robust against the class imbalance limitations, which is a common problem in clinical setting.
2. We present an in-depth analysis of existing AL strategies for multi-modal task – clinical report generation.

2 Related Work

This section provides the background of our proposed AL strategy BiCAL. We first formalize the active learning problem under the image-to-text generation task and set up the notation for the rest of the paper. Given a model $\mathcal{M}$, unlabelled image data pool X_{pool}. We denote an unlabelled input image as $x \in X_{pool}$, and the labelled text report as $y \in Y$, where $y = (y^1, ..., y^n)$, n is the number of tokens in the generated report. We define the labelled data pool X_{label} to contain image-report pairs, where $X_{\text{label}} \cap X_{\text{pool}} = \emptyset$. The whole data pool is $X_{all} := X_{label} \cup X_{pool}$. The model is parameterized by vector w, as follows:

$$\mathcal{M} = p_w(y \mid x) = p_w(y^1, ..., y^n \mid x) \tag{1}$$

An acquisition function representing the query heuristic in the AL setting is denoted as $a(x, \mathcal{M})$. At each active learning iteration, we acquire the label of a batch Q of b number of unlabelled instances from X_{pool} and add to the labelled data pool X_{label} using $a(x, \mathcal{M})$. The updated labelled data pool X_{label} is used to train the underlying model every iteration. This process iterates until a predefined budget $\mathcal{B}$ is depleted. Sampling from the pool is determined by the acquisition function as follows :

$$x^* = \text{argmax}_{x \in X_{\text{pool}}} a(x, \mathcal{M}) \tag{2}$$

2.1 Uncertainty-based and Diversity-based Active Learning

Uncertainty-based AL strategies often use a heuristic that can measure the model's uncertainty toward unlabelled data and choose the unlabelled data with the highest uncertainty (Lewis, 1995; Wang et al., 2019; Shannon, 2001). Gal et al. (2017) demonstrated the idea of measuring model uncertainty by combining Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011) with Bayesian formulation of Neural Networks such as Bayesian by Backprop (Blundell et al., 2015) and MC dropout (Gal and Ghahramani, 2016). However, uncertainty-based active learning typically depends on the underlying training model's predictions for uncertainty measurements. This dependence results in the "cold-start" problem (Yuan et al., 2020; Ash and Adams, 2020), where these methods are ineffective early in training due to the initial model's naivety.

On the other side, diversity-based Active Learning aims to select a subset of the data that can best represent the whole dataset, such that the model achieves similar performance to full-tuning when trained on the selected subset. There has been much previous work in this stream of designing AL strategies (Kim et al., 2006; Citovsky et al., 2021; Sener and Savarese, 2018).

2.2 Hybrid Active Learning

There have also been some hybrid AL methods that combine diversity and uncertainty in their design (Ash et al., 2019; Yuan et al., 2020). Approaches that infuse reinforcement learning into AL strategies which learn the selection heuristic from scratch were also seen (Fang et al., 2017; Liu et al., 2018; Vu et al., 2019). There has been close work on active learning to ours in natural language generation and abstractive text summarization, however, they focused on the single modal generation task (Tsvigun et al., 2023; Gidiotis and Tsoumakas, 2021a; Perlitz et al., 2023; Gidiotis and Tsoumakas, 2021b).

The closest work to ours is Contrastive Active Learning (CAL) proposed by (Margatina et al., 2021b). They hypothesized that if two data points are close in the underlying model feature space but result in very different underlying model predictive likelihood, then they may be lying on the model's decision boundary and therefore are a good candidate to query. CAL uses K-Nearest Neighbors (KNN) (Cover and Hart, 1967) to find and record the top k neighbouring points by their model representation encodings from the input. Then it computes the KL divergence (Kullback and Leibler, 1951) between the model's output probability of each unlabelled instance with their recorded k neighbours. The contrastive score of each unlabelled instance is then calculated by the average of all KL-divergence values of the neighbours. Ultimately, the data point with the highest contrastive score is selected to be queried.

3 Bidirectional Contrastive Active Learning

We identify the following limitations that existing AL methods have when training models in a clinical setting. In clinical settings, data for healthy or common sicknesses is often seen, while unhealthy or rare sicknesses are rare in the population, leading to an imbalanced dataset. This leads to models trained on such datasets that can converge easily on the common cases, and have poor performance on rare but important cases. Previous AL methods have not yet addressed this problem, as they are not able to explicitly identify important cases within the dataset automatically. The original CAL would identify two data points are neighbours if two data points have the same sickness, and if the model predicts differently for the two data points, they are considered as 'contrastive' and queried. Such a heuristic cannot locate the positive (unhealthy) cases efficiently, because negative (healthy) neighbour pairs would outweigh the positive (unhealthy) neighbour pair in the population, leading to the sampling process suffering from class imbalance and queries too many negative instances. Therefore, models trained using CAL achieve a bad performance in clinical efficacy and recalling positive cases, as revealed by our experiments in Table 1.

3.1 BiCAL Algorithm

BiCAL is robust to class imbalance datasets by its design and can automatically select rare but valuable cases within a dataset for the model to learn. This is done by bi-directionally augmenting the contrastive definition and measuring the contrastiveness in pre-trained embedding space, empowering the algorithm to select rare samples in domain datasets inherently.

We redefine two types of contrastive samples. For BiCAL, contrastive examples have to satisfy one of the following definitions:

1. Two data points with **similar** pre-trained embeddings but **different** pre-trained embeddings of their model generation outputs.
2. Two data points with **different** pre-trained embeddings but **similar** pre-trained embeddings of their model generation outputs.

The intuition behind the second augmented definition is that common cases and rare cases will most likely have the most different representations of each other within the dataset. Therefore, if a model generates similar outputs for two data points that have different representations, this means it is highly possible that at least one rare sample is within the two data points, and the current model hasn't trained enough on at least one of the two data points. Hence by augmenting the contrastive definition in BiCAL, we have increased the chance of querying a rare case, compared to CAL. Moreover, by leveraging pre-trained encoders, we isolate the underlying model in generating the uncertainty measure, alleviating the Cold-Start Learning problem (Yuan et al., 2020; Ash and Adams, 2020) – in the initial stage of training, the underlying model is naive due to the absence of domain knowledge, if we use the underlying model's encoder to generate uncertainty measure it would result in a decrease in the ineffectiveness of such uncertainty-based AL strategies.

Formally, each data point x_i should obtain k number of similar neighbours X_{close} and k number of dissimilar neighbours X_{far}.

$$\begin{aligned} X_{close} &:= \quad f(\Phi(x_i), \Phi(x_j)) < \epsilon \\ X_{far} &:= \quad f(\Phi(x_i), \Phi(x_j)) > \gamma \end{aligned} \tag{3}$$

For the first contrastive sample, the data point should satisfy the following condition:

$$f(\Omega(\mathcal{M}(x_i)), \Omega(\mathcal{M}(x^m_{close}))) > \gamma \tag{4}$$

For the second contrastive sample, the data point should satisfy the following conditions:

$$f(\Omega(\mathcal{M}(x_i)), \Omega(\mathcal{M}(x^m_{far}))) < \epsilon \tag{5}$$

Where $\Phi(.) \in \mathbb{R}^{d'}$ is a selected pre-trained image encoder that maps input x_i and x_j to its feature space. $\Omega(.) \in \mathbb{R}^{d''}$ is the selected pre-trained text encoder that maps the predicted output of underlying

Algorithm 1 Single iteration of BiCAL

```
Input: all data $X_{all}$, unlabeled data $X_{pool}$, acquisition size $b$, model $\mathcal{M}$, number of neighbours $k$, distance metric function $f(.)$, pre-trained image (encoding) function $\Phi(.)$, pre-trained text (encoding) function $\Omega(.)$, contrastive ratio $c \in [0,1]$, Total number of unlabelled data $N$, .
1  $S_{close} := \emptyset$ ; $S_{far} := \emptyset$
2  for $i$ in $1, \ldots, N$ do
3      $d_j \leftarrow f\big(\Phi(x_i), \Phi(x_j)\big)$                                  ▷ $x_j \in X_{all}, j = 1, ..., N$
4      $X_{close} \leftarrow$ Select k number of x$\in X_{all}$ with lowest d$_j$           ▷ $X_{close} = \{x^1_{close}, ..., x^k_{close}\}$; $j \neq i$
5      $X_{far} \leftarrow$ Select k number of x$\in X_{all}$ with highest d$_j$            ▷ $X_{far} = \{x^1_{far}, ..., x^k_{far}\}$
6      $\hat{Y}_{close} \leftarrow \mathcal{M}(X_{close})$
7      $\hat{Y}_{far} \leftarrow \mathcal{M}(X_{far})$
8      $\hat{y}_i \leftarrow \mathcal{M}(x_i)$
9      $s^i_{close} \leftarrow \frac{1}{k} \sum_{m=1}^{k} f\big(\Omega(\hat{y}_i), \Omega(\hat{y}^m_{close})\big)$
10     $s^i_{far} \leftarrow \frac{1}{k} \sum_{m=1}^{k} f\big(\Omega(\hat{y}_i), \Omega(\hat{y}^m_{far})\big)$
11     $S_{close} := S_{close} \cup \{s^i_{close}\}$ ; $S_{far} := S_{far} \cup \{s^i_{far}\}$
12 end
13 $Q_1 \leftarrow$ Select $b \times c$ number of x $\in X_{pool}$ with the highest $s_{close}$          ▷ $s_{close} \in S_{close}$
14 $Q_2 \leftarrow$ Select $b \times (1-c)$ number of x $\in X_{pool}$ with the lowest $s_{far}$        ▷ $s_{far} \in S_{far}$
Output: $Q_1 \cup Q_2$
```

model $\hat{y}_i$ to its feature space. $f(.)$ is a distance metric, such as Euclidean distance or cosine similarity. ϵ and γ represent the threshold for a very small and a very large distance value respectively, although in practice we adopt ranking instead of using a threshold. $\mathcal{M}(.)$ is the underlying training model of the active learning loop, such that $\hat{y}_i \leftarrow \mathcal{M}(x_i)$. We detail the single iteration of BiCAL's algorithm as follows:

Compute Neighbours We use the encoding function from the pre-trained model $\Phi(.)$ to map all the data points to its pre-trained embedding space. For each unlabelled instance x_i, we use cosine similarity $f(.)$ to measure the distances between the embeddings of x_i and all the other data points in the X_{all} (line 3). We record x_i's nearest (top k) and furthest (bottom k) neighbours in the embedding space by the distance calculated (lines 4-5).

Compute Contrastive Scores The unlabelled instance x_i and all its neighbours X_{close} and X_{far} will be passed to the underlying model $\mathcal{M}$ to generate their text outputs $\hat{y}$ (lines 6-8). The generated text from the model is then encoded by the selected pre-trained language model $\Omega(.)$ to obtain text embedding of the generated text. Using these embeddings, we can calculate two different contrastive scores for the unlabelled instance x_i (lines 9-10). The first contrastive score s^i_{close} is calculated by the average distance between the embedding of generated output of the unlabelled instances with their nearest neighbours, and the second one s^i_{far} is calculated with its furthest neighbours.

Query Two Contrastive Batches For each unlabelled instance x_i, we obtain two lists of contrastive scores S_{close} and S_{far}. We select the unlabelled instances using the two contrastive scores separately. For S_{close}, we select the top $b \times c$ number of instances, where b is the total intended batch size for query, and c is a hyperparameter "contrastive ratio" that controls the ratios of samples sampled from the two contrastive definitions. This gives us a batch of instances Q_1 of the first contrastive definition (line 13). For S_{far}, we select the bottom $b \times (1 - c)$ number of instances. This gives us a batch of instances Q_2 of the second contrastive definition (line 12). Ultimately, two batches Q_1 and Q_2 combines to give the output of BiCAL.

4 Experiment Settings

We assess BiCAL and other established AL methods' performance in training general multi-modal models to specify on the task of clinical report generation from chest X-ray images. In every active learning loop, the underlying model denoted as $\mathcal{M}$, was fine-tuned twice on the labelled pool X_{label}. Subsequently, we evaluated the model on the test dataset using various NLG metrics. Each experiment was run in 3 folds with different random seeds, each fold containing 10 active learning iterations, where 100 data points were queried per iteration, i.e. 1000 data points were queried in total. This choice of 1000 data points reflects real-world scenarios where active learning is applied when labelled data is not available. Our goal was to examine the efficiency and performance of active learning methods under constrained labelling budgets in a medical setting. In real-world AL scenarios, labelling a large size of unlabelled data is often impractical due to the significant expertise labelling effort required. Therefore, 1000 data points were deemed sufficient to assess the performance of the AL methods in our focus while mimicking a real-world AL situation. Future work could explore varying the number of training examples (e.g., 1500, 2000) to understand further the impact of labelled data quantity on active learning strategies in training medical models.

4.1 Baselines

We evaluate our proposed BiCAL against various literature Active Learning strategies:

1. **Random Sampling (RS):** Unlabelled instances are drawn at random.

2. **Normalized Sequence Probability (NSP):** Uses the probability of the generated sequence by the model as a measure of uncertainty.

$$\mathcal{NSP} = 1 - \exp\left\{\frac{1}{n}\sum_{i=1}^{n} log\mathbb{P}(y^i \mid y^1 \ldots, y^n, x)\right\}$$

   (Tsvigun et al., 2023; Wang et al., 2019).

3. **Expected Normalised Sentence Probability (ENSP):** Bayesian AL method where it has the same intuition as NSP.

$$\text{ENSP} = 1 - \mathbb{E}_{w\sim q_{\hat{\theta}}}\bar{p}_w(y|x)$$

   (Tsvigun et al., 2023; Ueffing and Ney, 2007; Wang et al., 2019).

4. **Expected Normalised Sentence Variance (ENSV):** Similar to ENSP but uses variance instead of expectation between the sequence probability.

$$\text{ENSV} = Var_{w\sim q_{\theta}}\bar{p}_w(y|x)$$

   (Tsvigun et al., 2023; Ueffing and Ney, 2007; Wang et al., 2019).

5. **Contrastive Active Learning (CAL):** SOTA AL method described in section 3 (Margatina et al., 2021b).

In addition, for **BiCAL**, we implemented two variants by varying the choice of pre-trained image encoder $\Phi(.)$ in the BiCAL algorithm. We have experimented with two types of pre-trained models, Dinov2 and CheSS, to examine the effect of different types of pre-trained image encoders in our algorithm. Dinov2 is an image model that is pre-trained on a general image dataset (Oquab et al., 2023), whereas CheSS is pre-trained on a CXR dataset (Cho et al., 2023). For the pre-trained text encoder $\Omega(.)$, we have fixed the selection to GatorTron (Yang et al., 2022) based on its SOTA performance in clinical NLP tasks (that outperforms BioBERT (Lee et al., 2019), ClinicalBERT (Huang et al., 2020), BioMegatron (Shin et al., 2020)).

4.2 Datasets

We used the labelled datasets MIMIC-CXR (Johnson et al., 2019a) and IU X-Ray (Demner-Fushman et al., 2015) for our simulation of active learning conditions. The IU X-Ray dataset contains 3,955 radiology reports with 7,470 associated chest X-ray images, while MIMIC-CXR includes 227,835 radiology reports with 377,110 associated chest X-ray images. Following the methodology from Chen et al. (2022), we excluded samples without accompanying reports. We partitioned the IU X-Ray dataset into training and testing sets using an 85%:15% ratio and used the official train-test split for MIMIC-CXR.

In our simulated active learning experiments, we queried only 1,000 data points. As it was impractical in terms of running time to run the experiment on the entire MIMIC-CXR dataset of 377,110 images, we leveraged the structured labels from MIMIC-CXR-JPG (Johnson et al., 2019b) and conducted stratified sampling to obtain a 10% subset of the training split (34,463 data points). This ensured that the subset closely mirrored the label distribution of the full MIMIC-CXR dataset. We used this stratified subset for training and the official test set for evaluation. We release the processed reports with their image IDs for both datasets in CSV files in the repository and provide the data distribution of MIMIC-CXR before and after subset sampling in Table 5 and 6 in the Appendix.

4.3 Setup

Experiments were conducted on a single NVIDIA RTX6000 GPU. We adopted the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 3e-5 and a weight decay of 3e-7. A warm-up scheduler was applied to the learning rate for the initial 200 steps. Due to computational constraints, we used a training batch size of 8 and limited the maximum number of tokens for generation to 100.

In our experiment, we fine-tuned a vision encoder-decoder model initialized with pre-trained Vision

	Precision	Recall	F-1 Score	Amount of training data
RS	0.450	0.252	0.168	1000
NSP	0.436	0.241	0.194	1000
ENSP	0.558	0.266	0.200	1000
ENSV	0.451	0.268	0.195	1000
CAL	0.326	0.221	0.187	1000
BiCAL Dinov2	0.403	0.255	0.191	1000
BiCAL CheSS	0.429	0.274	0.219	1000
Full Tune	0.309	0.273	0.259	34,463 (full subset)
R2Gen	0.333	0.274	0.276	377,110 (full data)
CCR	0.586	0.237	<0.300*	377,110 (full data)

Table 1: Clinical Efficacy Metrics across AL Strategies after 1000 data queried on MIMIC-CXR Dataset. * stared entries are estimated as the result is not found in the original paper. The best results over AL strategies of each metric are highlighted in blue. The performance of all models is averaged from three runs.

Methods	Example Outputs	BLEU 1-4
Reference	The NG tube extends inferiorly beyond the diaphragm into the fundus of the stomach. Again seen is moderate cardiomegaly. The pulmonary vascular congestion is stable. There are no new focal consolidations. The fissural loculation of pleural fluid along the left chest wall has not changed compared to the prior exam. There is no pneumothorax.	–
RS	The lungs are well expanded and clear. There is no pleural effusion or pneumothorax. The cardiomediastinal silhouette is within normal limits. There is no acute osseous abnormality.	[0.42, 0.36, 0.30, 0.24]
NSP	The heart is normal in size. The mediastinal and hilar contours appear within normal limits. The pulmonary vasculature is normal. There is no pleural effusion or pneumothorax. The lungs appear clear. There is no pulmonary edema.	[0.60, 0.53, 0.43, 0.36]
CAL	The patient is status post median sternotomy, CABG, and mitral valve replacement. The right-sided Port-A-Cath terminates in the low SVC. The right-sided pacemaker terminates in the low SVC. The mediastinal and hilar contours are unchanged. There is no pleural effusion or pneumothorax. The cardiac silhouette is top normal. There is no evidence of pneumothorax.	[0.85, 0.71, 0.56, 0.44]
BiCAL CheSS	The lungs are clear without focal consolidation, effusion, or pneumothorax. The cardiac and mediastinal silhouettes are within normal limits. No acute osseous abnormalities.	[0.37, 0.31, 0.24, 0.19]

Table 2: Case study of Generation Result on Negative Cases using Different AL Methods.

Transformers (ViT) (Dosovitskiy et al., 2020) and GPT-2 (Radford et al., 2019). These models were chosen for their popularity and strong performance in computer vision and natural language processing, respectively. Our primary focus was to investigate active learning strategies in a multi-modal task, so we did not explore other model choices. We utilized HuggingFace (Wolf et al., 2020) and Deepspeed (Rasley et al., 2020) to facilitate our experiment setup.

5 Results and Analysis

We used two types of evaluation metrics: natural language generation (NLG) metrics and domain-specific (clinical efficacy) metrics. This provided a comprehensive evaluation of the generated reports in terms of general and domain-specific performance. For NLG metrics, we reported BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores at each active learning iteration. For clinical efficacy metrics, we used the CheXpert (Irvin et al., 2019) model to label the generated and reference reports. We reported precision, recall, and F1 scores for the labeled categories of the generated and reference reports. This evaluation approach is widely used in chest X-ray clinical report generation tasks (Chen et al., 2022; Liu et al., 2019, 2021).

5.1 Clinical Efficacy Metrics

We first assessed the baseline methods and our strategy after 1000 queries on MIMIC-CXR using domain-specific metrics to examine the performance of active learning (AL) strategies, which is crucial for training clinical models. Table 1 displays the clinical efficacy metrics of various AL strategies based on 1000 data queries from a MIMIC-CXR dataset subset. The table's last three rows show the performance of our underlying model after fine-tuning for 10 epochs on the full MIMIC-CXR dataset subset, R2Gen (Chen et al., 2022), and the model (CCR) from Liu et al. (2019). These latter two are fully supervised models trained on the full MIMIC-CXR dataset, designed to excel in chest radiology report generation tasks, with their performance referenced directly from their published papers.

A notable observation is that BiCAL CheSS surpassed baseline methods in recall and F1 scores while maintaining a competitive average precision score. This suggests that the BiCAL CheSS approach effectively recognizes more rare cases (un-

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
CAL	0.4978	0.4177	0.3313	0.2685	0.3115	0.0996	0.2143
RS	0.4487	0.3762	0.3008	0.2456	0.3040	0.0979	0.2138
NSP	0.4832	0.3997	0.3160	0.2563	0.2994	0.1026	0.2178
ENSP	0.4238	0.3569	0.2868	0.2355	0.3066	0.1013	0.2205
ENSV	0.3588	0.3060	0.2477	0.2047	0.2939	0.0969	0.2119
BiCAL Dinov2	0.5025	0.4200	0.3343	0.2726	0.3096	0.1001	0.2183
BiCAL CheSS	0.3930	0.3299	0.2636	0.2153	0.2870	0.0905	0.2078

Table 3: Average NLG performance of different AL strategies after 1000 queries on MIMIC-CXR

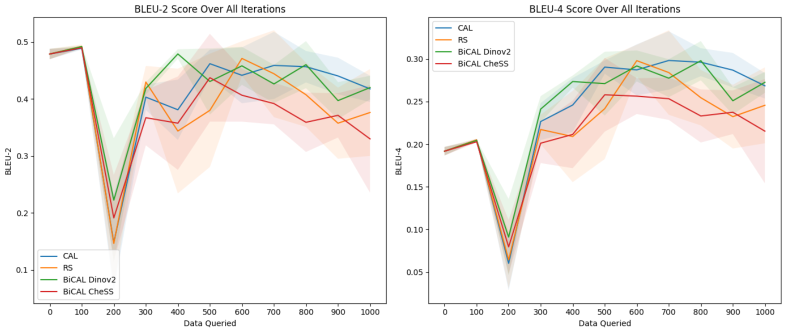

Figure 2: Average NLG Performance of AL Strategies and Best-performing Baselines on MIMIC-CXR

healthy scenarios) than other AL strategies, though it may occasionally increase false positives, as indicated by the precision score. In medical diagnostics, catching every potential disease case (reducing false negatives) is crucial. Therefore, high recall is preferable to high precision, making BiCAL's performance desirable in our context and demonstrating BiCAL CheSS's superiority in generating clinically accurate reports.

Remarkably, the BiCAL CheSS method achieved a recall score that surpasses models fine-tuned on the entire MIMIC-CXR subset (Full Tune). Additionally, it achieved competitive performance with fully supervised models R2Gen and CCR, with a better recall score and an F1 score not much lower. This is noteworthy, considering this performance was achieved with only 1000 data points (less than 0.3% of the whole MIMIC-CXR).

An interesting observation is that although CAL performed well in the NLG metrics on the MIMIC-CXR dataset (Figure 3), its clinical precision and recall scores were the least impressive among all methods. This suggests that while CAL trains models to produce seemingly accurate reports, these might not be clinically sound. By augmenting the contrastive bidirectionally and utilizing pre-trained encoders, the domain-specific performance of this contrastive active learning approach is largely enhanced, demonstrating the success of our approach. We include a case study of generation performance on rare cases using various AL methods in Table 7 in the Appendix.

Furthermore, evidence of the task's complexity is seen in the last three rows of Table 1. These rows include results from R2Gen and CCR, models specifically tailored for chest X-ray report generation and comprehensively trained on the full MIMIC-CXR dataset. Despite their specialized design, their clinical performance remains relatively low. This underscores the inherent challenge of our downstream task—clinical report generation. The intricacies in medical images may be difficult for the underlying model's capability to learn, suggesting that to truly elevate clinical accuracy, superior clinical models adept at the task may need to be designed.

5.2 Natural Language Generation Metrics

We found that for the IU X-ray dataset, no single strategy consistently outperformed the others. Notably, RS and NSP showed marginally better performance during the initial four iterations in both BLEU and ROUGE metrics. For the MIMIC-CXR dataset, CAL performed slightly better in ROUGE scores, while BiCAL was competitive with CAL in BLEU scores, as shown in Figure 3.

The varying performance of CAL across the MIMIC-CXR and IU X-ray datasets suggests that CAL's superiority did not extend to the IU X-ray dataset. This may be due to the different data volumes. Smaller datasets result in a limited unlabeled data pool, potentially narrowing batch sample variance and minimizing observable performance variance. Consequently, the queried batches of different AL strategies on the IU X-ray dataset have more overlap than on MIMIC-CXR, leading to similar performance across strategies.

For the BLEU score, BiCAL Dinov2 performed better than all strategies before 500 queries but was surpassed by CAL afterwards (≥ 500), though it remained competitive. For ROUGE scores, CAL consistently retained slightly better performance starting from 300 queried data. This comparison demonstrates BiCAL's competitiveness in NLG metrics. As shown in Table 3, after 1000 queries, BiCAL Dinov2 achieved the best performance in all BLEU scores and the second-best performance in all ROUGE scores.

Although BiCAL only surpassed other literature AL methods in some NLG metrics, it remained competitive with the best-performing baseline methods. However, language models have been criticized for producing hallucinated text (Ouyang et al., 2022; Stiennon et al., 2020; Ziegler et al., 2019). In a medical setting, our priority is creating accurate clinical reports, not just authoritative-sounding ones. We believe the relatively worse performance of BiCAL CheSS is due to the hallucination problem of LLMs.

CAL and other methods suffer from class imbalance data and may select more healthy cases for training, leading to hallucinated models, that are good at generating good negative (healthy) reports containing many common phrases. In contrast, BiCAL may have a higher proportion of positive cases, training a model with higher clinical efficacy. However, this model's ability to write comprehensive healthy reports that match the reference deteriorates. This results in worse performance on NLG metrics due to the class imbalance problem (more negative cases than positive in the test set, causing the model to generate negative reports more often). This hypothesis is supported by our analysis of the generation results of the models under different active learning methods, including a case study in Table 2. It can be seen that although all reports are saying the candidate contains no significant diseases, but other methods learn to give a more comprehensive healthy report, which results in a higher BLEU score. Thus due to the imbalanced dataset problem, the average NLG score of the other methods may exceed BiCAL despite being less clinically accurate in positive cases (shown in clinical efficacy metrics in Table 1). We also include a positive case study from our analysis to show BiCAL's ability to train clinically accurate models in Table 7.

c	Precision	Recall	F-1 Score
0	0.381	0.254	0.177
0.25	0.376	0.241	0.170
0.50	0.430	0.274	0.219
0.75	0.516	0.250	0.188
1	0.417	0.264	0.199

Table 4: Micro Average of Precision, Recall, and F-1 Score on CheXpert classification Result of BiCAL using different contrastive ratio c after 1000 data queried on MIMIC-CXR Dataset

5.3 Ablation Study

In Sections 5.1 and 5.2, we discussed the impact of different image encoders on the BiCAL algorithm, comparing those pre-trained on a general image dataset (Dinov2) and a Chest X-ray dataset (CheSS). Additionally, a crucial component of the BiCAL algorithm is the contrastive ratio, denoted as c, which determines the sampling ratio between two contrastive definitions in a batch. Our previous experiments used a default c value of 0.5, meaning an equal split between the two contrastive definitions. As shown in Table 4, for clinical efficacy metrics, BiCAL performs best when c is 0.5 in terms of clinical recall and F1 scores. For clinical precision, a c value of 0.75 seems optimal. The poorest performance in terms of clinical recall is observed at $c = 0.25$. This suggests that while a c value of 0.5 may not be the best for NLG metrics, it ensures the generation of higher clinical quality reports by achieving the best recall of diseases in the generated reports.

6 Conclusion

In this work, we present a study on the effectiveness of current active learning methods for domain-specific multi-modal learning, specifically on the task of clinical report generation from chest X-ray images. We identified the challenge of class imbalance in domain-specific active learning and addressed it by introducing BiCAL, a new active learning technique. BiCAL excelled in both NLG and domain-specific (clinical efficacy) metrics, notably outperforming baselines in clinical recall and F1-score.

We found that existing AL strategies demonstrate similar performance in NLG metrics for the task of clinical report generation from chest X-ray images. This may be due to the complexity of our task, which requires training the model to acquire clinical expertise to generate accurate and clinically sound reports. Interestingly, our tests revealed that an AL strategy's high performance in NLG metrics does not ensure equal success in domain-specific (clinical) performance, possibly due to the hallucination properties of language models. We hope this work provides valuable insights and can act as a starting point for researchers in the future on the task of active learning in multi-modal clinical tasks.

Ethical Consideration and Limitations

We note that despite the success of BiCAL in our study of clinical report generation, in practice, its performance is yet to be confirmed. We have simulated our experiments based on a labelled dataset where the radiology report was collected under a monitored condition such that their format may achieve a certain level of consistency (Johnson et al., 2019a; Demner-Fushman et al., 2015). However, in practice, the queried data's label report may vary based on different radiologist labellers, which may cause noise in the training dataset, which may affect the effectiveness of BiCAL.

We identify that for this work have used sensitive personal data that is related to the health sector. We used MIMIC-CXR (Johnson et al., 2019a) and IU X-Ray (Demner-Fushman et al., 2015) datasets in this project. We note that both datasets have been de-identified, where they have removed all personal health information (PHI). This has ensured the privacy and confidentiality of the individuals. During this project, we handled the data responsibility and used it only for the purpose of research. No attempt at re-identification of the datasets is made. We have also signed the data use agreement for MIMIC-CXR before we use the data. We note that MIMIC-CXR and IU X-rays, just like all datasets, may contain inherent biases based on patient information such as where the data is collected. Moreover, active learning is a technique that samples data based on a certain heuristic, which therefore may introduce additional bias in the sampling and training of the model. This work researches the effectiveness of active learning in clinical report generation, we recognize this potential bias that may be introduced by our research, and this also comes along with our work's contribution to the improvement of the field of active learning in the clinical sector.

Due to the difficulties in acquiring publicly available domain-specific image-report pair data, we chose to work with the task of clinical report generation from chest X-rays. As we designed the BiCAL algorithm, it was not tailored to the clinical report generation task that we conducted our experiments. Moreover, we believe that with the intricacies and high level of expertise required in the medical domain, we believe experiments conducted in this domain can provide valuable insight and act as a good reference for AL's performance on domain-specific learning in general. However, further work should be done in the future to test the performance of BiCAL in other domains, given that the data is available.

References

Jordan T. Ash and Ryan P. Adams. 2020. On warm-starting neural network training. *Preprint*, arXiv:1910.08475.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. 2023. Sequential modeling enables scalable learning for large vision models. *Preprint*, arXiv:2312.00785.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks. *Preprint*, arXiv:1505.05424.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2022. Generating radiology reports via memory-driven transformer. *Preprint*, arXiv:2010.16056.

Kyungjin Cho, Ki Duk Kim, Yujin Nam, Jiheon Jeong, Jeeyoung Kim, Changyong Choi, Soyoung Lee, Jun Soo Lee, Seoyeon Woo, Gil-Sun Hong, Joon Beom Seo, and Namkug Kim. 2023. CheSS: Chest x-ray pre-trained model via self-supervised contrastive learning. *Journal of Digital Imaging*.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. 2021. Batch active learning at scale. *Preprint*, arXiv:2107.14263.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Dina Demner-Fushman, Marc Kohli, Marc Rosenman, Sonya Shooshan, Laritza Rodriguez, Sameer Antani, George Thoma, and Clement Mcdonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for bert: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. *Preprint*, arXiv:1708.02383.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR.

Alexios Gidiotis and Grigorios Tsoumakas. 2021a. Bayesian active summarization. *Preprint*, arXiv:2110.04480.

Alexios Gidiotis and Grigorios Tsoumakas. 2021b. Uncertainty-aware abstractive summarization. *ArXiv*, abs/2105.10155.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint*, arXiv:1904.05342.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Preprint*, arXiv:1901.07031.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019a. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019b. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Seokhwan Kim, Yu Song, Kyungduk Kim, Jeong-Won Cha, and Gary Geunbae Lee. 2006. MMR-based active machine learning for bio named entity recognition. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 69–72, New York City, USA. Association for Computational Linguistics.

Solomon Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

David D Lewis. 1995. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive attention for automatic chest x-ray report generation. In *Findings*.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. *Preprint*, arXiv:1904.02633.

Ming Liu, Wray L. Buntine, and Gholamreza Haffari. 2018. Learning how to actively learn: A deep imitation learning approach. In *Annual Meeting of the Association for Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Katerina Margatina, Loïc Barrault, and Nikolaos Aletras. 2021a. On the importance of effectively adapting pretrained language models for active learning. In *Annual Meeting of the Association for Computational Linguistics*.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021b. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. 2023. Active learning for natural language generation. *Preprint*, arXiv:2305.15040.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. *Preprint*, arXiv:1708.00489.

Burr Settles. 2009. Active learning literature survey.

Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates. *Preprint*, arXiv:2101.08133.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: Larger biomedical domain language model. *Preprint*, arXiv:2010.06060.

Emma Slade and Kim M. Branson. 2022. Deep reinforced active learning for multi-class image classification. *Preprint*, arXiv:2206.13391.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Rinyoichi Takezoe, Xu Liu, Shunan Mao, Marco Tianyu Chen, Zhanpeng Feng, Shiliang Zhang, and Xiaoyu Wang. 2023. Deep active learning for computer vision: Past and future. *APSIPA Transactions on Signal and Information Processing*, 12(1).

Akim Tsvigun, Ivan Lysenko, Danila Sedashov, Ivan Lazichny, Eldar Damirov, Vladimir Karlov, Artemy Belousov, Leonid Sanochkin, Maxim Panov, Alexander Panchenko, Mikhail Burtsev, and Artem Shelmanov. 2023. Active learning for abstractive text summarization. *Preprint*, arXiv:2301.03252.

Nicola Ueffing and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40.

Thuy-Trang Vu, Ming Liu, Dinh Q. Phung, and Gholamreza Haffari. 2019. Learning how to active learn by dreaming. In *Annual Meeting of the Association for Computational Linguistics*.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 791–802, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *Preprint*, arXiv:2203.03540.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. *arXiv preprint arXiv:2010.09535*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

7 Appendix

Table 5: Label Distribution for Full MIMIC-CXR Dataset

	-1.0	0.0	1.0	N/A
Atelectasis	4.53%	0.67%	20.11%	74.69%
Cardiomegaly	2.65%	6.98%	19.68%	70.68%
Consolidation	1.90%	3.50%	4.73%	89.87%
Edema	5.78%	11.25%	11.86%	71.10%
Enlarged Cardiomediastinum	4.11%	2.32%	3.15%	90.42%
Fracture	0.24%	0.39%	1.93%	97.44%
Lung Lesion	0.50%	0.38%	2.76%	96.36%
Lung Opacity	1.68%	1.35%	22.62%	74.36%
No Finding	0.00%	0.00%	33.12%	66.88%
Pleural Effusion	2.55%	11.92%	23.83%	61.69%
Pleural Other	0.34%	0.06%	0.88%	98.73%
Pneumonia	8.03%	10.68%	7.27%	74.02%
Pneumothorax	0.50%	18.59%	4.55%	76.36%
Support Devices	0.10%	1.53%	29.21%	69.15%

Table 6: Label Distribution for Stratified Subset of MIMIC-CXR Dataset

	-1.0	0.0	1.0	N/A
Atelectasis	4.62%	0.72%	19.94%	74.72%
Cardiomegaly	2.62%	6.83%	19.82%	70.73%
Consolidation	1.83%	3.52%	4.62%	90.03%
Edema	5.79%	11.53%	11.51%	71.17%
Enlarged Cardiomediastinum	4.06%	2.29%	3.10%	90.55%
Fracture	0.24%	0.38%	1.93%	97.45%
Lung Lesion	0.55%	0.42%	2.64%	96.38%
Lung Opacity	1.68%	1.40%	22.71%	74.21%
No Finding	0.00%	0.00%	33.26%	66.74%
Pleural Effusion	2.57%	11.99%	23.54%	61.90%
Pleural Other	0.32%	0.06%	0.87%	98.75%
Pneumonia	8.09%	10.56%	7.39%	73.97%
Pneumothorax	0.50%	18.36%	4.65%	76.48%
Support Devices	0.09%	1.48%	29.43%	69.00%

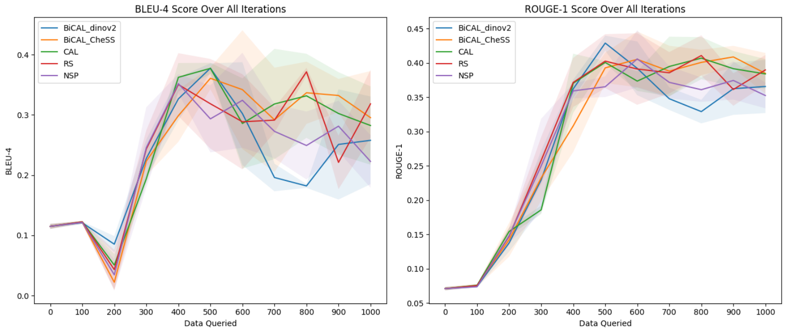

Figure 3: Average NLG Performance of AL Strategies and Best-performing Baselines on IU X-ray

Methods	Example Outputs
Reference	Lung volumes are low. Mild to moderate enlargement cardiac silhouette is unchanged, accentuated by the presence of low lung volumes. The aorta remains tortuous. Mediastinal and hilar contours are stable. There is continued mild pulmonary vascular congestion without overt pulmonary edema. Patchy and linear opacities in the lung bases likely reflect areas of atelectasis. No pneumothorax or pleural effusion is clearly evident. Percutaneous gastrostomy catheter is incompletely imaged.
RS	The lungs are clear. There is no pleural effusion or pneumothorax. Cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities.
NSP	The heart is normal in size. The mediastinal and hilar contours appear within normal limits. There is no pneumothorax. The pulmonary vasculature is normal. There is no pleural effusion or pneumothorax. There is no pneumomediastinum.
CAL	The heart is mildly enlarged. There is mild prominence of pulmonary vascularity with mild interstitial edema. There is no pleural effusion or pneumothorax. The mediastinal and hilar contours are unremarkable. There is no evidence of pneumomediastinum.
BiCAL CheSS	The cardiac silhouette is mildly enlarged. The aorta is tortuous. There is mild cardiomegaly. There is no pleural effusion or pneumothorax. The hilar contours are normal. There is mild pulmonary vascular congestion.

Table 7: Case study of Generation Result on Positive Cases using Different AL Methods. Green: The generated diagnosis is matched with reference. Red: The generated diagnosis is incorrect compared to the reference. Yellow: The generated diagnosis is not mentioned in the reference.

Disease	**RS**	**NSP**	**ENSP**	**ENSV**	**CAL**	**BiCAL Dinov2**	**BiCAL CheSS**	**Full Tune**
No Finding	0.0738	0.0799	0.0766	0.0843	0.0911	0.0750	0.1068	0.1507
Enlarged Cardiomediastinum	0.2183	0.2410	0.2378	0.2333	0.2462	0.2318	0.2386	0.2958
Cardiomegaly	0.2475	0.2592	0.1783	0.2354	0.2781	0.1829	0.4177	0.5113
Lung Lesion	1.0000	1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.0000
Lung Opacity	1.0000	0.6667	0.6667	1.0000	0.4333	0.3333	0.5000	0.3798
Edema	0.1781	0.1869	0.1555	0.1548	0.1757	0.1669	0.1584	0.2315
Consolidation	0.2879	0.4248	0.3455	0.3292	0.3029	0.3241	0.2981	0.3160
Pneumonia	0.2000	0.1221	1.0000	0.1176	0.0870	0.0000	0.1481	0.0887
Atelectasis	0.3846	0.3509	0.3636	0.3333	0.2773	0.5000	0.3333	0.2739
Pneumothorax	0.5621	0.6102	0.5876	0.5701	0.5569	0.5713	0.5917	0.5949
Pleural Effusion	0.4567	0.5131	0.4949	0.4945	0.4558	0.4906	0.4876	0.6016
Pleural Other	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
Fracture	0.0000	0.0000	1.0000	1.0000	0.0323	0.0000	0.0000	0.1667
Support Devices	0.6939	0.6418	0.6986	0.7545	0.6253	0.7610	0.7282	0.7096
Macro Average	0.4502	0.4355	0.5575	0.4505	0.3258	0.4026	0.4292	0.3086

Table 8: Precision on CheXpert classification Result between reference and generated report across AL Strategies after 1000 queries on MIMIC-CXR

Disease	**RS**	**NSP**	**ENSP**	**ENSV**	**CAL**	**BiCAL Dinov2**	**BiCAL CheSS**	**Full Tune**
No Finding	0.9042	0.8314	0.9042	0.8391	0.7011	0.7893	0.6590	0.7356
Enlarged Cardiomediastinum	0.4196	0.3924	0.4030	0.3970	0.3587	0.4267	0.4237	0.3869
Cardiomegaly	0.1221	0.1512	0.1042	0.1753	0.2267	0.1945	0.4083	0.3757
Lung Lesion	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Lung Opacity	0.0005	0.0010	0.0010	0.0010	0.0199	0.0005	0.0005	0.1921
Edema	0.1357	0.1614	0.1801	0.1376	0.0752	0.1402	0.1961	0.1145
Consolidation	0.6344	0.1336	0.4760	0.5180	0.2395	0.4812	0.5120	0.2372
Pneumonia	0.0022	0.0229	0.0000	0.0022	0.0131	0.0000	0.0218	0.0196
Atelectasis	0.0041	0.0164	0.0296	0.0008	0.0961	0.0041	0.0008	0.0895
Pneumothorax	0.7024	0.8285	0.7880	0.8968	0.7810	0.7900	0.8297	0.5608
Pleural Effusion	0.5581	0.5395	0.5318	0.6064	0.4310	0.5322	0.5302	0.6205
Pleural Other	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Fracture	0.0000	0.0000	0.0000	0.0000	0.0034	0.0000	0.0000	0.0034
Support Devices	0.0400	0.2928	0.3039	0.1717	0.1452	0.2040	0.2551	0.4797
Macro Average	0.2517	0.2408	0.2658	0.2676	0.2208	0.2545	0.2741	0.2725

Table 9: Recall on CheXpert classification Result between reference and generated report across AL Strategies after 1000 queries on MIMIC-CXR

Disease	**RS**	**NSP**	**ENSP**	**ENSV**	**CAL**	**BiCAL Dinov2**	**BiCAL CheSS**	**Full Tune**
No Finding	0.1365	0.1458	0.1412	0.1531	0.1612	0.1370	0.1838	0.2502
Enlarged Cardiomediastinum	0.2872	0.2986	0.2991	0.2939	0.2920	0.3004	0.3053	0.3353
Cardiomegaly	0.1635	0.1910	0.1315	0.2010	0.2498	0.1886	0.4129	0.4331
Lung Lesion	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Lung Opacity	0.0010	0.0020	0.0020	0.0020	0.0381	0.0010	0.0010	0.2552
Edema	0.1540	0.1732	0.1669	0.1457	0.1054	0.1524	0.1753	0.1532
Consolidation	0.3961	0.2033	0.4004	0.4026	0.2675	0.3873	0.3768	0.2710
Pneumonia	0.0043	0.0385	0.0000	0.0043	0.0227	0.0000	0.0380	0.0321
Atelectasis	0.0081	0.0314	0.0547	0.0016	0.1427	0.0081	0.0016	0.1349
Pneumothorax	0.6245	0.7028	0.6732	0.6971	0.6502	0.6631	0.6907	0.5773
Pleural Effusion	0.5023	0.5260	0.5127	0.5448	0.4431	0.5105	0.5080	0.6109
Pleural Other	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Fracture	0.0000	0.0000	0.0000	0.0000	0.0062	0.0000	0.0000	0.0067
Support Devices	0.0756	0.4021	0.4236	0.2797	0.2357	0.3217	0.3779	0.5724
Macro Average	0.1681	0.1939	0.2004	0.1947	0.1868	0.1907	0.2194	0.2594

Table 10: F1 Score on CheXpert classification Result between reference and generated report across AL Strategies after 1000 queries on MIMIC-CXR