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

Anonymous ACL submission

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, such as in the medical sector. However, training these models requires vast amounts of labelled data, a challenge in medicine 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, such as generating clinical reports from images. In this work, we proposed a novel AL strategy, **Bi**directional **C**ontrastive **A**ctive **L**earning strategy (BiCAL), that uses both image and text latent spaces to identify contrastive samples to selects batch to query for labels. BiCAL is robust to cold-start learning problem in AL and class imbalance data by its design. Our experiments show that BiCAL outperforms standard methods in clinical efficacy metrics, improving recall by 2.4% and F1 score by 9.5%, showcasing its effectiveness in actively training clinical 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 a model train (Settles, 2009). This technique allows the model to achieve optimal performance while lowering the cost of annotation. It has shown great potential in the field of NLP recently by reducing the volume of annotated data while not sacrificing model performance (Shelmanov et al., 2021; Dor et al., 2020; Shen et al., 2017; Margatina et al., 2022).

However, relatively few have explored the application of active learning to fine-tune multi-modal models in image-to-text downstream tasks. There has been close work of AL on natural language generation (NLG) (Gidiotis and Tsoumakas, 2021a; Tsvigun et al., 2023; Perlitz et al., 2023) and neural machine translation (NMT) (Haffari et al., 2009; Ambati et al., 2010). However, in specific domains like the clinical sector, obtaining quality labelled data is challenging due to the clinical expertise required for accurate annotation (Budd et al., 2021; Chen et al., 2015), which is costly in both time and money. This motivates us to explore active learning's application in the clinical report generation tasks. Moreover, in domain-specific active learning, there exists two challenges: 1) Cold-start learning: Uncertainty-based AL strategies usually rely on the underlying training model to provide a measure of uncertainty. They became ineffective since the underlying training model does not acquire domain-specific knowledge in the early training phase (Yuan et al., 2020; Ash and Adams, 2020). 2) Class imbalance in datasets like medical ones, where existing AL methods struggle to prioritize positive (unhealthy) samples in its selection, hindering model learning for positive cases – our prime interests in training medical models.

In this study, we introduce a novel AL strategy BiCAL that is tailored to address the two challenges 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 develop a novel AL strategy BiCAL that is robust to the cold start learning and class imbalance problem in domain-specific active learning.

2. We show that BiCAL outperforms the literature in clinical efficacy metrics while maintaining competitive in NLG metrics.

3. We present an in-depth analysis of existing AL strategies for clinical report generation. To the

best of our knowledge, this is the first study of AL for image-to-text generation tasks.

2 Related Work

This section provides the background of our proposed AL strategy BiCAL.

2.1 Clinical Report Generation

General image-captioning by large deep learning models has been seen successful in application (Li et al., 2022, 2023; Radford et al., 2021; Dosovitskiy et al., 2020; Li et al., 2021). Motivated by this success, many have explored its potential on the radiology report generation task, where most adopt encoder-decoder architecture to achieve image-to-text translation. (Chen et al., 2022; Jing et al., 2017; Yuan et al., 2019; Li et al., 2018; Liu et al., 2019b; Li et al., 2019; You et al., 2022; Hou et al., 2021; Tanida et al., 2023). Moreover, to facilitate research in clinical report generation, researchers have also curated and released clinical image-report pair datasets. (Johnson et al., 2019a; Demner-Fushman et al., 2015; Irvin et al., 2019).

2.2 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).

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). When considering the closest literature, the work most aligned with ours are active learning in natural language generation and abstractive text summarization (Tsvigun et al., 2023; Gidiotis and Tsoumakas, 2021a; Perlitz et al., 2023; Gidiotis and Tsoumakas, 2021b). BiCAL differs from these methods in the way that it is able to select positive samples in a medical dataset inherently through contrastive sampling, leading to a better model that can achieve a higher recall of diseases.

3 Bidirectional Contrastive Active Learning

As this work focuses on investigating AL's application in clinical report generation, we formalize the active learning problem under this 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 $\mathcal{X}_{label}$ to contain image-report pairs. The whole data pool is $X_{all} := X_{label} \cup X_{pool}$. The model is parameterized by vector w, as follows:

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

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 :

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

3.1 Limitation of Contrastive Active Learning

Contrastive Actice Learning (CAL) (Margatina et al., 2021) hypothesize that if two data points are close in the model feature space but result in very different 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. However, CAL exhibits two crucial limitations in domain-specific active learning as mentioned earlier:

Cold Start Problem The standard CAL approach depends on the encoding function of the base training model. This leads to the "cold-start problem". At the beginning of training, the model may not possess domain-specific knowledge, hence the encoding of input data points by the underlying training model became uninformative, leading to inaccurate neighbours drawn by the KNN algorithm, hence making CAL ineffective.

Targeting Positive Cases Under the original CAL's contrastive definition, if there are two data points that have the same sickness, they first become neighbours of each other, and if the model predicts differently of the two data points, they are considered as 'contrastive' and queried. However, in medical datasets, it is very often seen that the dataset will have a class imbalance problem, where the proportion of negative (healthy) cases outweighs the positive (unhealthy) cases. CAL cannot locate the positive cases efficiently, because negative neighbours pairs would outweigh the positive neighbours pair in population, leading to the sampling process suffering from the class imbalance and queries too many negative instances. Therefore, models trained using CAL achieves a bad performance in clinical efficacy and recalling positive cases, as revealed by our experiments.

3.2 BiCAL Algorithm

The main improvements of BiCAL from CAL (Margatina et al., 2021) are summarised as follows. 1) We address the Cold-Start Learning problem by leveraging pre-trained encoders to reduce the algorithm's reliance on the underlying model $\mathcal{M}$ in providing embeddings of input data. 2) BiCAL inherently select positive (unhealthy) samples regardless of the class imbalance problem. This is done by augmenting the contrastive definition into bidirectional. The augmented definition combined with the quality embeddings from the pre-trained encoder empowers the algorithm to select positive samples.

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 negative cases and positive cases will most likely have the most different representations of each other. 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 positive sample is within the two data points. Hence by augmenting the contrastive definition in BiCAL, we have increased the chance of querying a positive case, compared to CAL.

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 condition:

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

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 $\quad d_j \leftarrow f\Big(\Phi(x_i), \Phi(x_j)\Big)$ $\quad \triangleright\ x_j \in X_{all}, j = 1, ..., N$
4 $\quad X_{close} \leftarrow$ Select k number of x$\in X_{all}$ with lowest d_j $\quad \triangleright\ X_{close} = \{x^1_{close}, ..., x^k_{close}\}; j \neq i$
5 $\quad X_{far} \leftarrow$ Select k number of x$\in X_{all}$ with highest d_j $\quad \triangleright\ X_{far} = \{x^1_{far}, ..., x^k_{far}\}$
6 $\quad \hat{Y}_{close} \leftarrow \mathcal{M}(X_{close})$
7 $\quad \hat{Y}_{far} \leftarrow \mathcal{M}(X_{far})$
8 $\quad \hat{y}_i \leftarrow \mathcal{M}(x_i)$
9 $\quad s^i_{close} \leftarrow \frac{1}{k} \sum_{m=1}^{k} f\Big(\Omega(\hat{y}_i), \Omega(\hat{y}^m_{close})\Big)$
10 $\quad s^i_{far} \leftarrow \frac{1}{k} \sum_{m=1}^{k} f\Big(\Omega(\hat{y}_i), \Omega(\hat{y}^m_{far})\Big)$
11 $\quad S_{close} := S_{close} \cup \{s^i_{close}\}$; $S_{far} := S_{far} \cup \{s^i_{far}\}$
12 **end**
13 $Q_1 \leftarrow$ Select $b \times c$ number of x $\in X_{pool}$ with the highest s_{close} $\quad \triangleright\ s_{close} \in S_{close}$
14 $Q_2 \leftarrow$ Select $b \times (1 - c)$ number of x $\in X_{pool}$ with the lowest s_{far} $\quad \triangleright\ s_{far} \in S_{far}$

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 point 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 "contratsive 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 evaluate BiCAL on image-to-text generation task in the medical domain, specifically, Chest X-ray clinical report generation task. 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.

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., 2021).

In addition, For **BiCAL**, we implemented two variants. BiCAL algorithm requires the specification of two pre-trained encoders, one for encoding image input, and one for encoding the generated text output of the underlying training model $\mathcal{M}$, denoted as $\Phi(.)$ and $\Omega(.)$ respectively. For the pre-trained image encoder $\Phi(.)$, we have experimented with two types of pre-trained model 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 dataset MIMIC-CXR (Johnson et al., 2019a) and IU X-Ray (Demner-Fushman et al., 2015) for a simulation of active learning conditions. IU X-ray contains a total of 3955 radiology reports with 7470 associated chest X-ray images, and MIMIC-CXR contains 227,835 radiology reports with 377,110 associated chest X-ray images. For both datasets, we adopted the methodology from Chen et al. (2022) to exclude samples without accompanying reports. The IU X-RAY dataset was partitioned into training and testing sets using a ratio of 85%:15%, while the MIMIC-CXR dataset was divided according to its official train-test split.

In our simulated active learning experiments, we only queried 1000 data points in total, therefore it was unnecessary and impractical in terms of time constraint, to run an active learning experiment on the full dataset of MIMIC-CXR, which consists of 377,110 images. To address this, we have leveraged the structured labels provided by MIMIC-CXR-JPG (Johnson et al., 2019b). We conducted stratified sampling to obtain a 10% subset of the train split of dataset (34463 data points). This approach ensured that the subset dataset closely mirrored the label distribution of the full dataset of MIMIC-CXR. The label distributions before and after this stratified sampling are depicted in appendices. Consequently, for MIMIC-CXR, we employed the stratified sampled subset for training and used the official test set for evaluation. We release the processed reports with their image ID for both datasets in CSV files in the repository.

4.3 Setup

Experiments are run on a single NVIDIA RTX6000 GPU. We adopted the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate set to 3e-5 and a weight decay of 3e-7. A warm-up scheduler was applied to the learning rate for the initial 500 steps. Due to computational constraints,

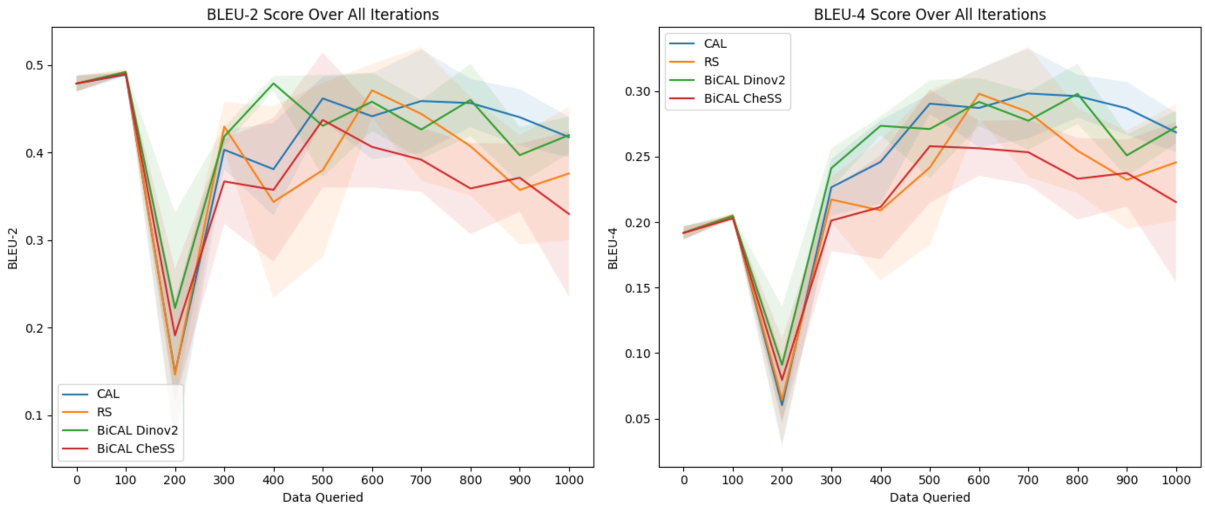

Figure 1: Average NLG Performance of AL Strategies and Best-performing Baselines on MIMIC-CXR

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

we used a training batch size of 8. We limited the maximum number of tokens for generation to 100.

In the experiment, we fine-tune a vision encoder-decoder model that is initialized by pre-trained vision transformers (ViT) (Dosovitskiy et al., 2020) and GPT-2 (Radford et al., 2019), the choice of the two models chosen is based on their popularity and good performance in CV and NLP field respectively, we do not delve into investigating different choice of this underlying model in this work as our primary focus is to investigate AL strategy in clinical report generation task. We utilized HuggingFace (Wolf et al., 2020) and Deepspeed (Rasley et al., 2020) to aid the set-up of our experiments.

We use two types of evaluation metrics, the traditional natural language generation (NLG) metrics and clinical efficacy metrics. For NLG metrics, we report BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores every active learning iteration. For clinical efficacy metrics, we use CheXpert (Irvin et al., 2019) to label the generated reports and the reference reports, and we report the precision, recall and F1 scores of the labelled category of the generated and reference reports, this approach is used widely in this task (Chen et al., 2022; Liu et al., 2019a, 2021).

5 Results and Analysis

5.1 Natural Language Generation Metrics

We found that for the IU X-ray dataset, no single strategy consistently surpasses the others. Notably, RS and NSP exhibit marginally better performance during the initial four iterations in both BLEU and ROUGE metrics. For the MIMIC-CXR dataset, we find that CAL performs slightly better than other strategies in ROUGE scores, BiCAL is able to achieve competitive performance with CAL in BLEU score as shown in Figure 1.

We observe a varying performance of CAL across MIMIC-CXR and IU X-Ray datasets, where the superiority of CAL doesn't show in the IU X-Ray dataset. This might stem from the different data volumes. Smaller datasets result in a limited unlabeled data pool, potentially narrowing batch sample

	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(Chen et al., 2022)	0.333	0.274	0.276	377,110 (full data)
CCR (Liu et al., 2019a)	0.586	0.237	<0.300*	377,110 (full data)

Table 2: Clinical Efficacy Metrics across AL Strategies after 1000 data queried on MIMIC-CXR Dataset. * stared entries is estimated as the result is not found in the original paper. The best results over AL strategies of each metric are highlighted in blue. Detailed results can be found in Appendix.

variance and consequently minimizing observable performance variance, i.e. the queried batch of different AL strategies on IU-dataset will have more overlap than on MIMIC-CXR, hence leading to similar performance using different strategies.

We can observe that for the BLEU score, BiCAL Dinov2 has a better performance than all strategies before 500 queries, but is surpassed by CAL afterwards (≥ 500) though it still remains competitive. For ROUGE scores, CAL consistently retains a slightly better performance starting from 300 queried data. This comparison result has demonstrated BiCAL's competitiveness in its performance on NLG metrics. On the other side, as shown in Table 1, after 1000 queries, BiCAL Dinov2 achieves the best performance in all BLEU scores, while able to achieves the second-best performance in all ROUGE scores.

In conclusion, for the NLG metrics, although BiCAL only surpasses literature AL methods in some metrics, it remains competitive with the best-performing baseline methods. However, it's worth noting that language models have faced criticism for producing text that might sound authoritative but can be misleading (Ouyang et al., 2022; Stiennon et al., 2020; Ziegler et al., 2019). In a medical setting, our priority is creating clinically accurate reports, instead of reports that are authoritative sounding. With this in mind, we'll further assess the baseline methods and our strategy after 1000 queries on MIMIC-CXR using the clinical efficacy metric.

5.2 Clinical Efficacy Metrics

Table 2 displays the clinical efficacy metrics of various active learning (AL) strategies, based on 1000 data queries on a MIMIC-CXR dataset subset. The table's last three rows display the performances of our underlying model after fine-tuning for 5 epochs on the full MIMIC-CXR dataset subset, R2Gen (Chen et al., 2022), and the model in paper (Liu et al., 2019a), respectively. The latter two are full supervision models where they were trained with full MIMIC-CXR and were designed to excel in chest radiology report generation task. Their performance was referenced directly from their published paper.

A notable observation is that BiCAL CheSS surpasses baseline methods in the recall and F-1 score while maintaining an average competitive precision score. This suggests that the BiCAL CheSS approach can effectively recognize a higher number of actual positive cases (unhealthy scenarios) than other AL strategies, although may occasionally lead to an increase in false positive errors, as indicated by the precision score. In the context of medical diagnostics, it's crucial to catch every potential disease case (reduce false negatives). This is because we do not want to miss any illness, meaning that high recall is preferable to high precision. Therefore BiCAL's performance is a desirable behaviour in our context and demonstrates BiCAL CheSS's superiority in generating better clinically accurate reports.

Remarkably, the BiCAL CheSS method achieves a recall score that surpasses the models that are fine-tuned on the entire subset of MIMIC-CXR (Full Tune). Moreover, it is able to achieve competitive performance with fully supervised model R2Gen and CCR, where it achieves a better recall score and a f1 score that is not lower by a large margin. We highlight this result, as we note that this performance is achieved only on 1000 data points (less than 0.3% of the whole MIMIC-CXR).

An interesting observation is that although CAL performs well in the NLG metrics on the MIMIC-CXR dataset (Figure 1), but its clinical precision and recall scores are the least impressive among all methods, not to mention in comparison with BiCAL Chess. This suggests that while CAL trains models to produce seemingly accurate reports, these might not be clinically sound. Also, it demonstrates that by augmenting the contrastive bidirectionally and utilizing pre-trained encoders, the clinical efficacy performance of this contrastive active learning approach can be largely enhanced, suggesting the successfulness of our approach.

Furthermore, evidence of the task's complexity is seen in the last three rows of Table 2. These rows include results from R2Gen and CCR, models specifically tailored for chest x-ray report generation and trained comprehensively on the full MIMIC-CXR dataset. Despite their specialized design, their clinical performance still is at a relatively low level. This observation underscores the inherent challenge of our downstream task - clinical report generation, this may be due to the intricacies in medical images are hard to learn by the underlying model's capability. To truly elevate clinical accuracy, there may be a need to design superior clinical models adept at the task. It's worth noting that the potential of active learning is inherently bounded by the capability of the base model. In essence, if a model's upper limit is, say, 90% accuracy, then even the most optimal active learning strategy would struggle to push its performance beyond this threshold.

5.3 Ablation Study

In the BiCAL algorithm, a crucial component is the contrastive ratio, denoted as c. This ratio determines how a batch of BiCAl is queried, defining the sampling ratio between two contrastive definitions. The previous experiments used a default c value of 0.5, meaning an equal split between the two contrastive definitions. In the section, we fix BiCAL to its CheSS variant version and varied c within the range [0,1] to explore its influence on BiCAL's performance.

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 3: 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

As shown in Table 3, for clinical efficacy metrics, the BiCAL performs best when c is 0.5 for clinical recall and F1 scores. Regarding clinical precision, the value of $c = 0.75$ seems optimal. The poorest performance in terms of clinical recall is observed at $c = 0.25$. This suggests that a c value of 0.5 might not be the best for NLG metrics, but it assures a model that can generate higher clinical quality reports as it achieves the best recalling of diseases in the generated report.

6 Conclusion

In this study, we evaluated the effectiveness of current active learning methods for generating clinical reports from chest X-ray images. We introduced BiCAL, a new active learning technique, which excelled in both NLG and clinical metrics, notably outperforming baselines in clinical recall and F1-score. We find that existing AL strategies demonstrate similar performance in NLG metrics. This may be due to the complexity of our task, which requires training the model to acquire domain-specific knowledge to generate clinical-sounding reports. A possible solution is the Actune framework: first fine-tuning the language decoder autoregressively on a medical text dataset, then actively learning on the downstream task (Yu et al., 2022). Interestingly, our tests revealed that an AL strategy's high performance in NLG metrics doesn't ensure equal success in clinical metrics, which may be due to untruthful generation by language models.

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, this 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.

References

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Jordan T. Ash and Ryan P. Adams. 2020. On warm-starting neural network training.

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*.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks.

Samuel Budd, Emma C Robinson, and Bernhard Kainz. 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

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

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.

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.

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.

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

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423, Boulder, Colorado. Association for Computational Linguistics.

Benjamin Hou, Georgios Kaissis, Ronald Summers, and Bernhard Kainz. 2021. Ratchet: Medical transformer for chest x-ray diagnosis and reporting.

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.

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.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2017. On the automatic generation of medical imaging reports. In *Annual Meeting of the Association for Computational Linguistics*.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *ArXiv*, abs/1805.08298.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. *ArXiv*, abs/1903.10122.

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. 2019a. Clinically accurate chest x-ray report generation.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019b. Clinically accurate chest x-ray report generation. *ArXiv*, abs/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.

Katerina Margatina, Loïc Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764*.

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.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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.

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.

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.

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.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

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.

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.

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.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2022. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. *ArXiv*, abs/2203.10095.

Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. 2022. AcTune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1422–1436, Seattle, United States. Association for Computational Linguistics.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. *ArXiv*, abs/1907.09085.

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*.

A Appendix

Table 4: 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 5: 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%

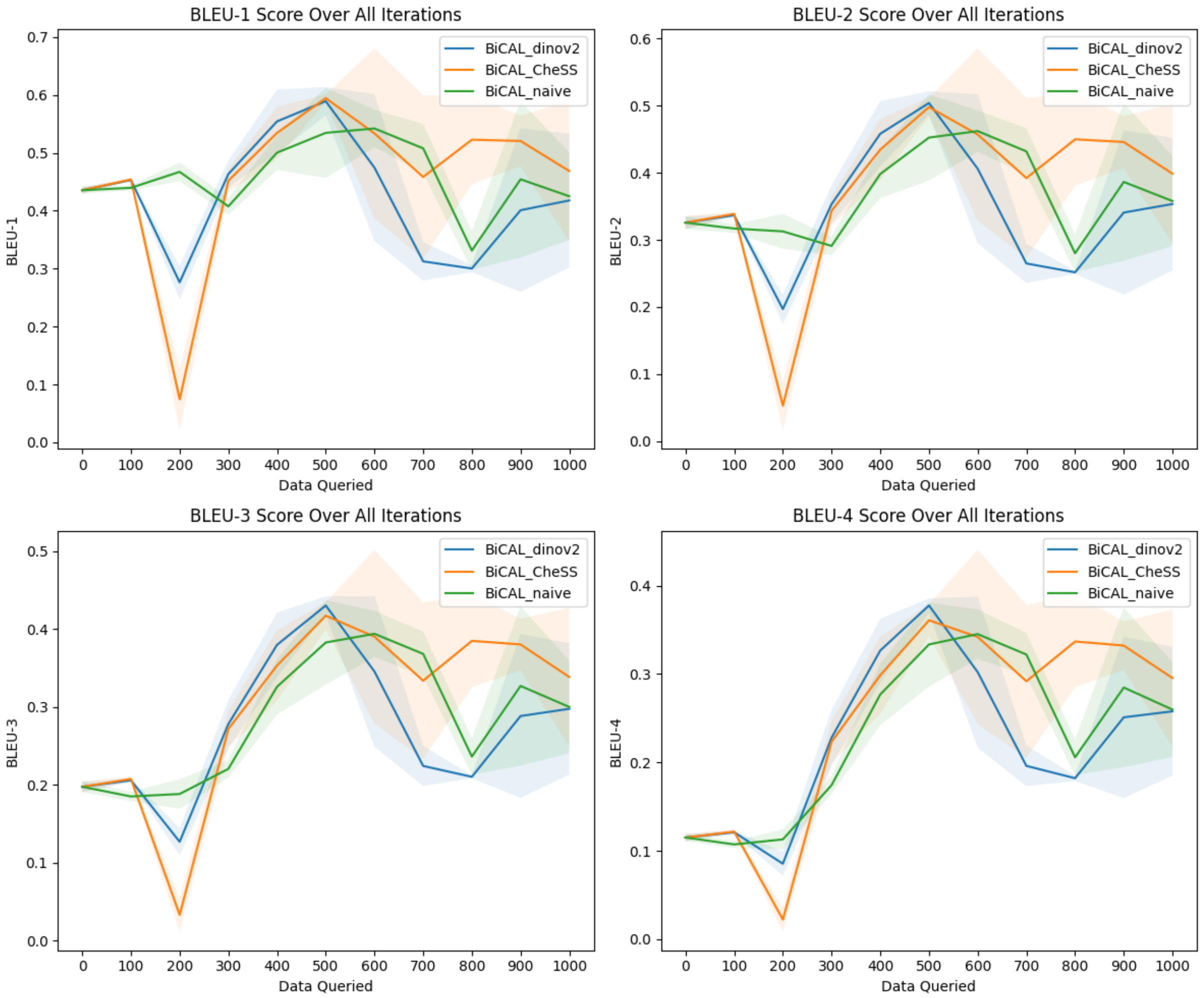

Figure 2: BLEU scores of BiCAL using Different Image Encoder on IU X-Ray dataset

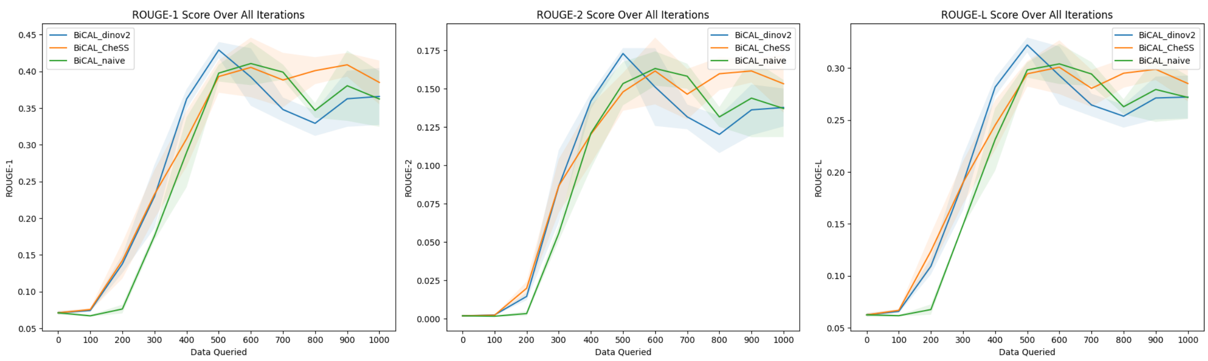

Figure 3: ROUGE scores of BiCAL using Different Image Encoder on MIMIC-CXR dataset

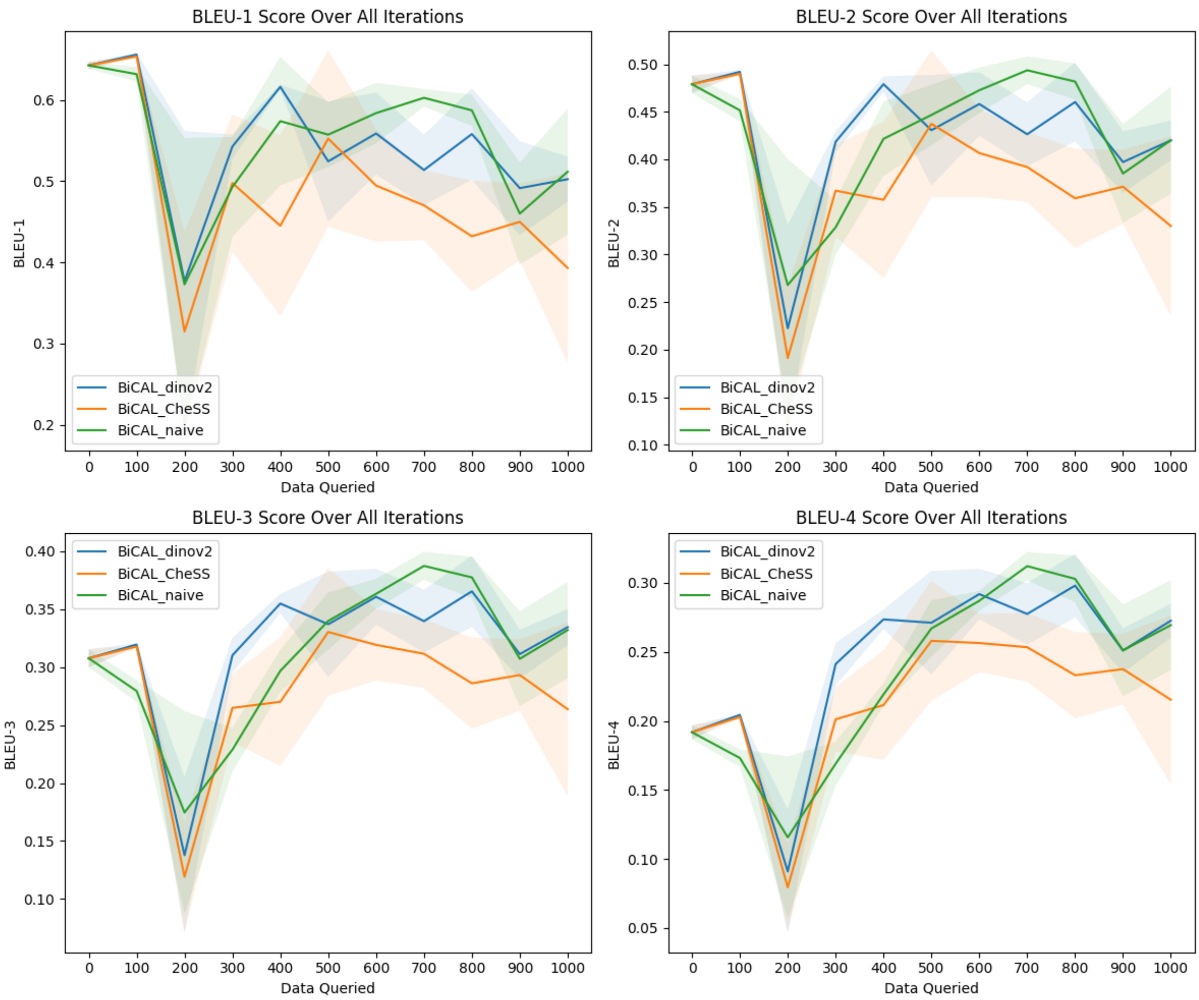

Figure 4: BLEU scores of BiCAL using Different Image Encoder on IU X-Ray dataset

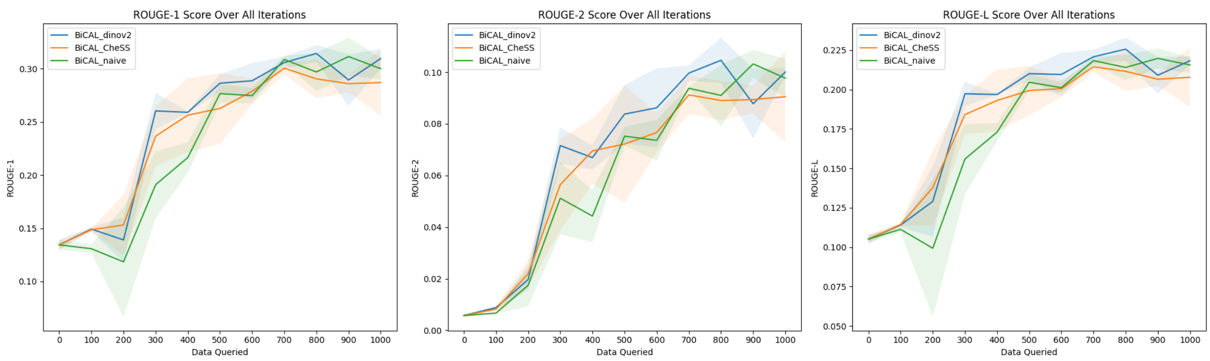

Figure 5: ROUGE scores of BiCAL using Different Image Encoder on MIMIC-CXR dataset

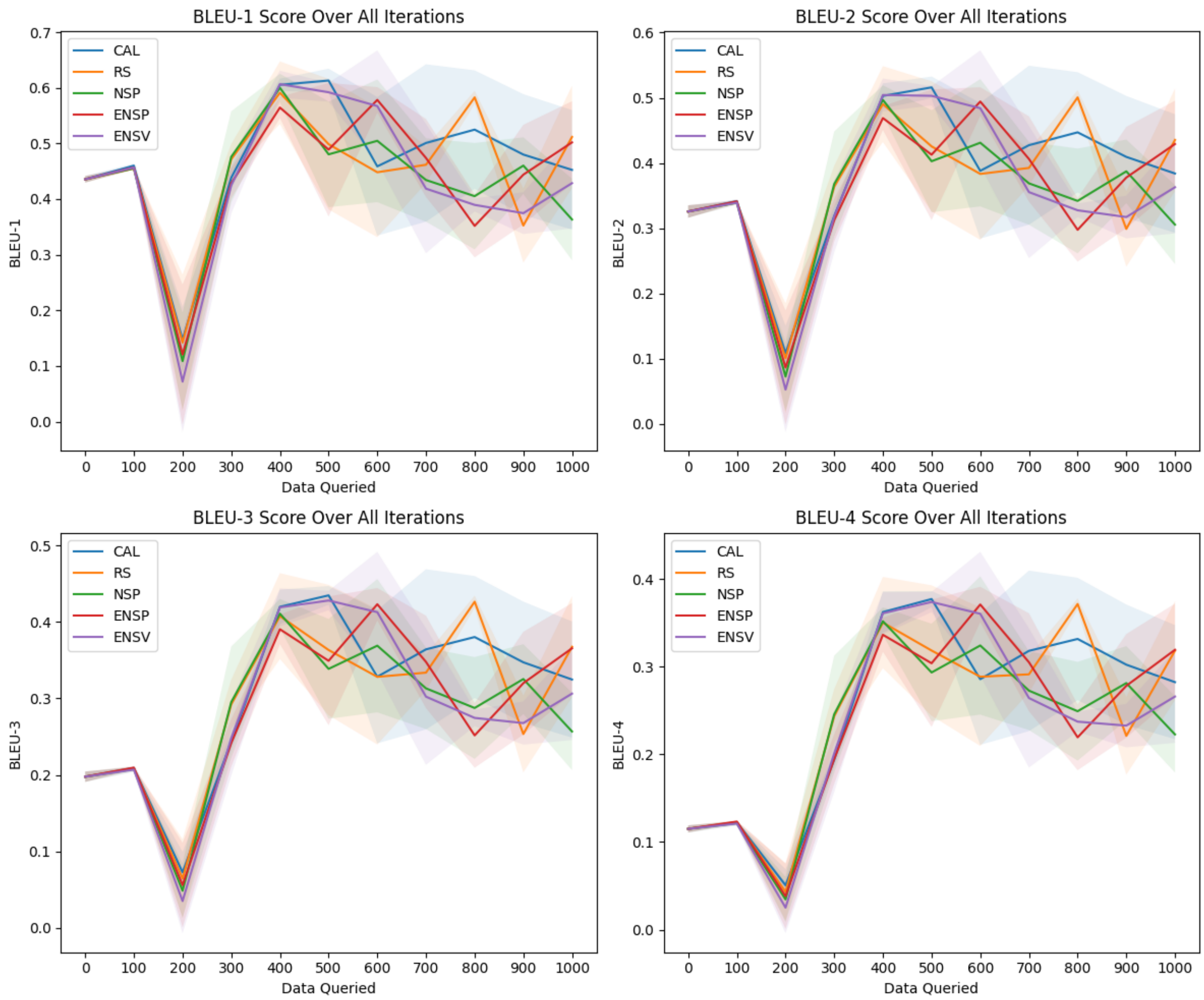

Figure 6: BLEU scores of Different Baseline AL Strategies on IU X-Ray dataset

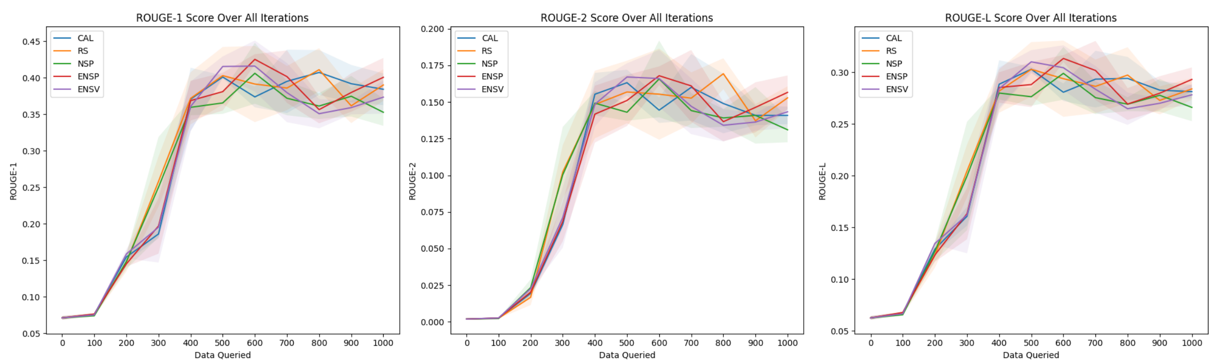

Figure 7: ROUGE scores of Different Baseline AL Strategies on MIMIC-CXR dataset

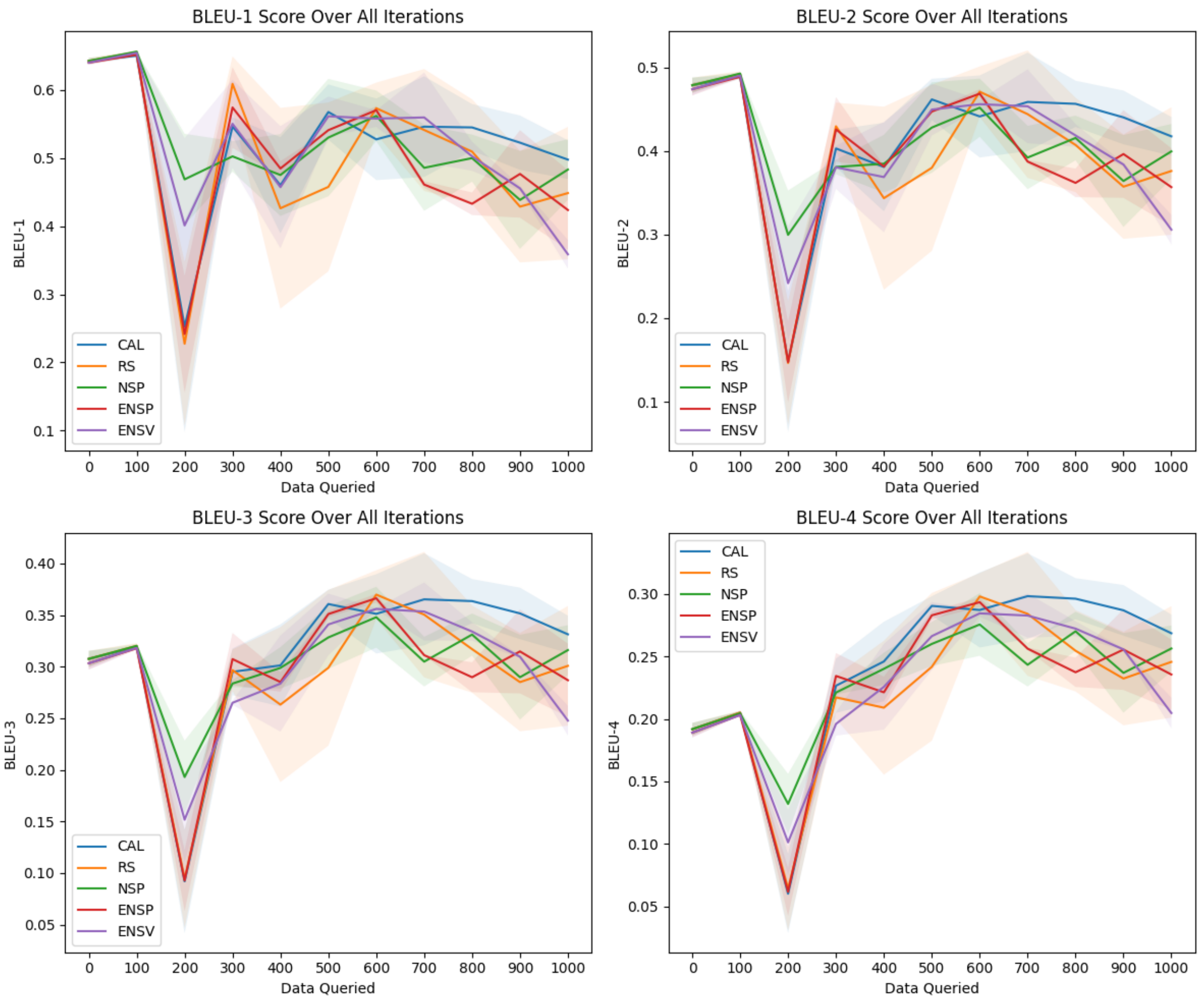

Figure 8: BLEU scores of Different Baseline AL Strategies on IU X-Ray dataset

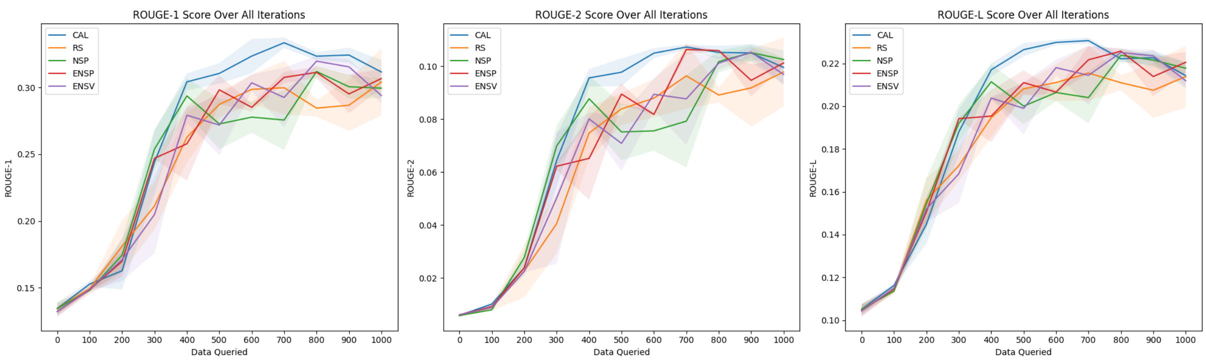

Figure 9: ROUGE scores of Different Baseline AL Strategies on MIMIC-CXR dataset

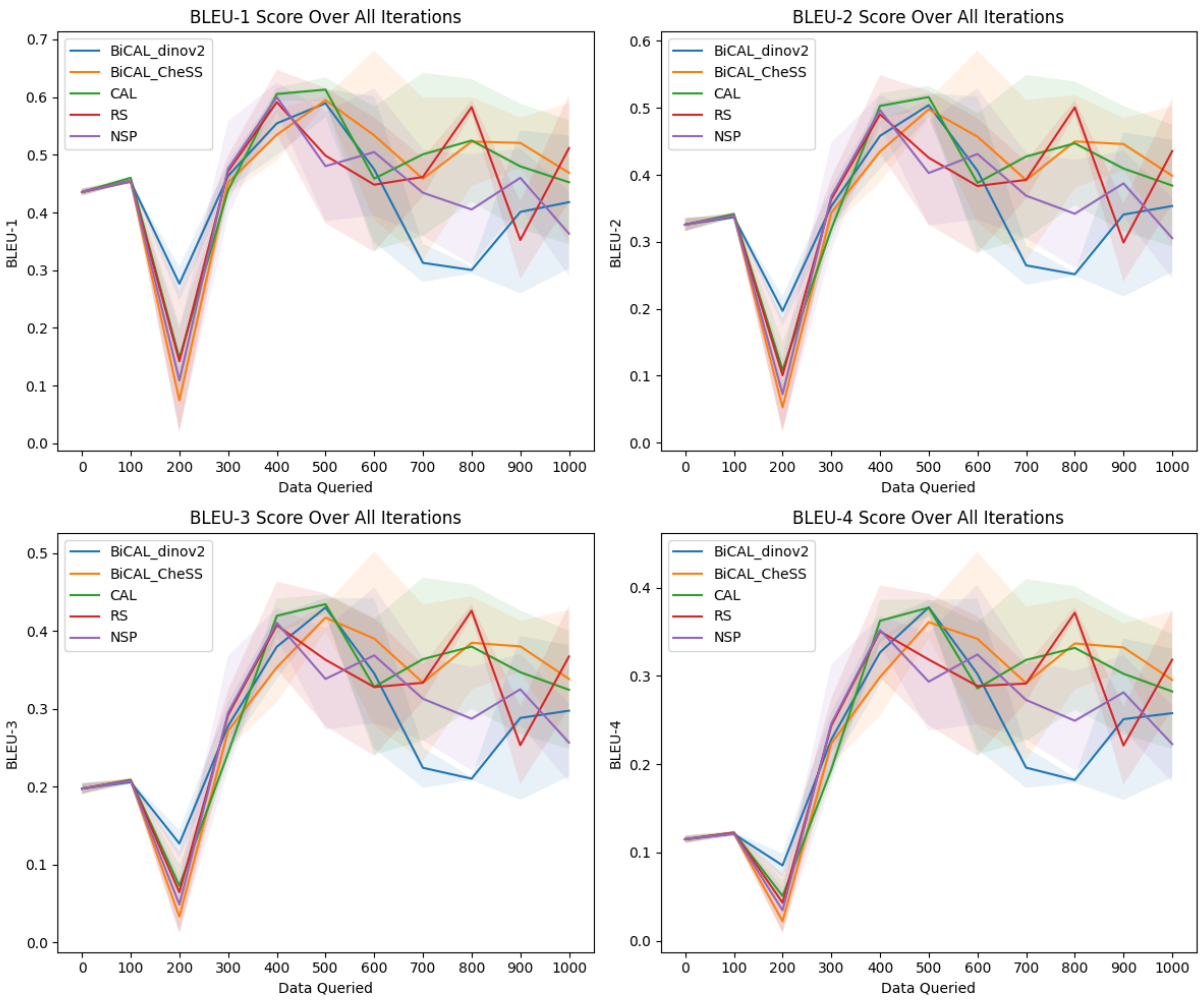

Figure 10: BLEU scores of BiCAL and Best Performing Baseline AL Strategies on IU X-Ray dataset

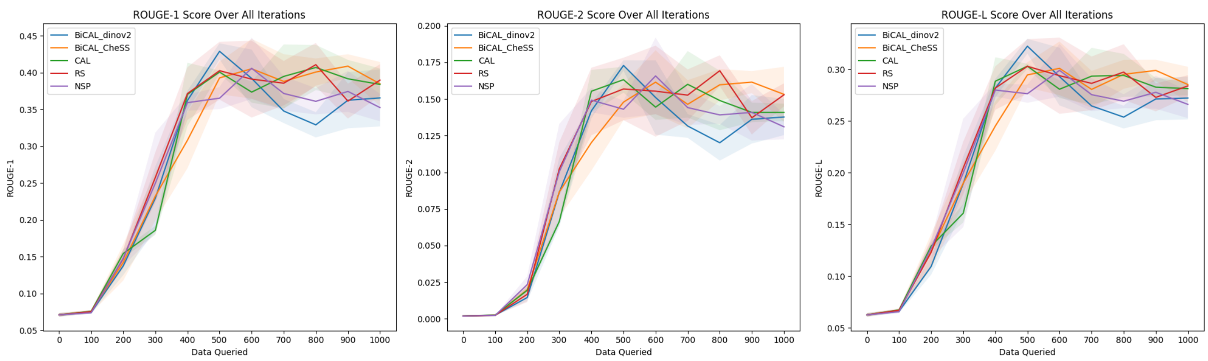

Figure 11: ROUGE scores of BiCAL and Best Performing Baseline AL Strategies on MIMIC-CXR dataset

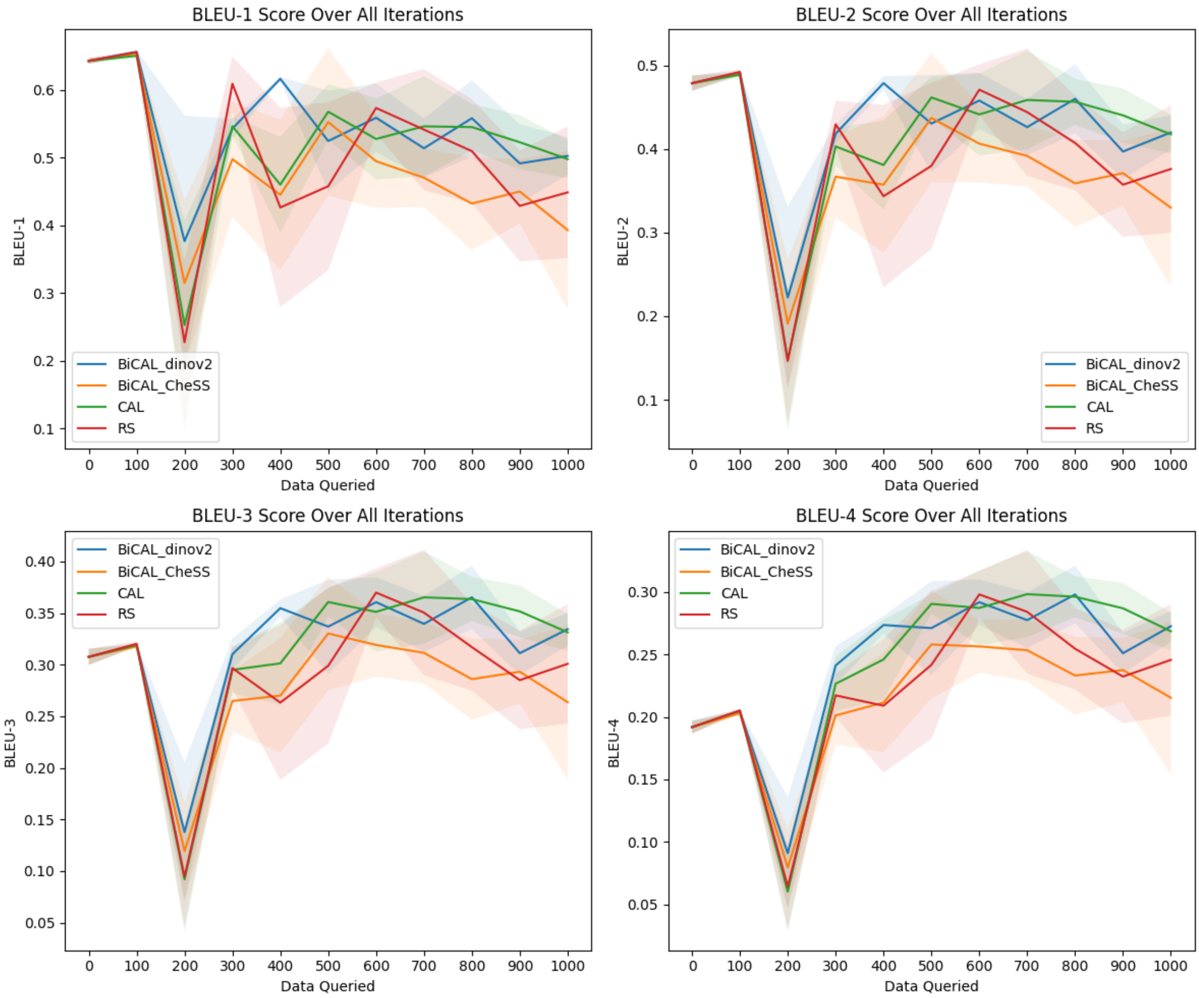

Figure 12: BLEU scores of BiCAL and Best Performing Baseline AL Strategies on IU X-Ray dataset

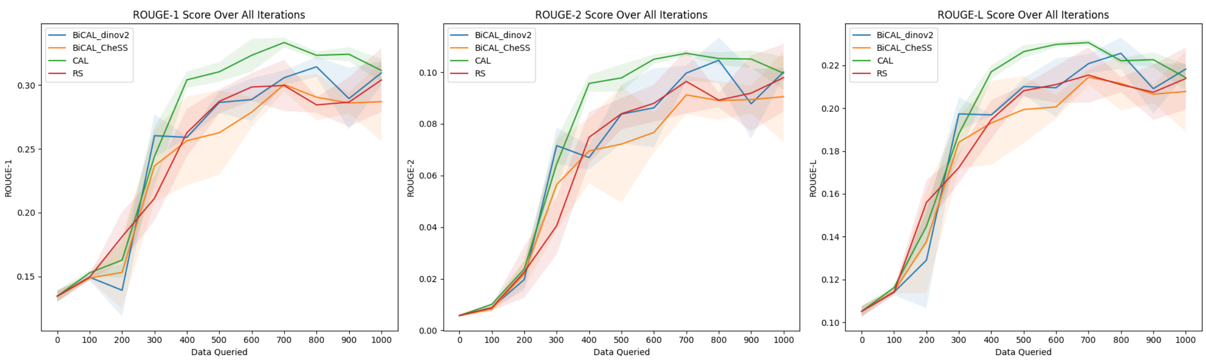

Figure 13: ROUGE scores of BiCAL and Best Performing Baseline AL Strategies on MIMIC-CXR dataset

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 6: 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 7: 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 8: F1 Score on CheXpert classification Result between reference and generated report across AL Strategies after 1000 queries on MIMIC-CXR

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
CAL	0.4524	0.3839	0.3246	0.2824	0.3841	0.1408	0.2812
RS	0.5116	0.4354	0.3674	0.3184	0.3900	0.1529	0.2838
NSP	0.3633	0.3055	0.2567	0.2228	0.3526	0.1310	0.2660
ENSP	0.5019	0.4292	0.3654	0.3193	0.4005	0.1566	0.2930
ENSV	0.4285	0.3628	0.3062	0.2660	0.3733	0.1433	0.2781
BiCAL naive	0.4251	0.3579	0.3001	0.2598	0.3624	0.1370	0.2717
BiCAL Dinov2	0.4179	0.3534	0.2978	0.2578	0.3658	0.1377	0.2721
BiCAL CheSS	0.4688	0.3986	0.3386	0.2954	0.3849	0.1532	0.2851

Table 9: Average NLG performance after 1000 queries on IU X-Ray

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 naive	0.5117	0.4201	0.3318	0.2694	0.3001	0.0977	0.2156
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 10: Average NLG performance after 1000 queries on MIMIC-CXR