Using LLM's for Semantic Text Identification in Medical Analysis

Mourad Zeynalov **Claire Liang** **Ankith Rajendran** **Ruchitha Kuthethoor**

Abstract

The goal of this project is to train a large language model on a limited dataset to perform semantic text classification for medical analysis, while considering key privacy concerns. We focus specifically on improving semantic matching of medical reports to diagnostic labels through the use of fine-tuned LLMs. The approach leverages fine-tuning a pre-trained LLM on a section of the procured dataset, then applying few-shot learning paradigms to further enhance the adaptability of the model to diverse medical cases. In addition, we experiment with lightweight fine-tuning techniques such as LoRA to address resource constraints and accelerate the process of supervised fine tuning.

1 Hypothesis

A large language model trained on a specialized dataset of medical reports will perform better at diagnosing disease labels present in medical reports than syntactic text similarity methods.

2 Introduction

The process of analyzing medical reports is a vital part of healthcare, especially in the diagnosis of certain disease labels based on the tests or scans undertaken. It is predominately a human-specific task, often performed by experienced medical practitioners. This leads to a bottleneck due to the limit of reports a single individual can read and draw observations from. To tackle this, there have been recent attempts to address this issue through NLP applications, primarily through tasks related to text summarization and machine translation. However, these techniques are limited in that the medical field is very domain specific, and procuring large amounts of data is hard due to patient sensitivity and issues with privacy. These summarized reports also do not entirely solve the problem, as the medical practitioner would still have to go through the reports to make a diagnosis. In cases where the summarized or translated reports are inaccurate or missing information, it can also prove to be fatal, since the medical field deals with human lives, and every single incorrect classification could be potentially life-threatening. Another problem is that there are complex relationships in medical reports that cannot be captured by lexical comparison methods, and hence finding the right metrics to evaluate these models is of paramount importance.

To address these limitations, we propose training a large language model on a limited dataset size, given the dearth of high-quality medical reports to generate class labels corresponding to diseases. For this project, the dataset of choice is the CheXpert Plus dataset, which consists of X-ray results and their corresponding reports. The language model is fed the medical reports as input, and then conditioned to generate a vector of 14 labels, 13 of those corresponding to different diseases and the last label being the "No Finding" label. These reports are also de-identified, meaning the anonymity of patients is ensured and thus no sensitive information is fed into the model. Different techniques will be implemented to test the effectiveness on the dataset, and a direct comparison between these methods will be analyzed to infer the best methods when dealing with medical data. The main focus of this project is to attempt to improve the semantic text classification ability of large language models when dealing with medical data, under consideration of privacy concerns. The methods explored in sequence are as follows: a baseline consisting of using cosine similarities between embeddings of reports to generate class labels, generation of labels using a base LLM, in this case the *Llama 3 8B* base model, and then finally an instruction tuned model using a split of the dataset for supervised fine tuning.

3 Related Work

Recent research by Xu et al. (2024) explores using large language models (LLMs) to enhance semantic analysis in specialized medical texts. Their framework employs GPT-4 for zero-shot text identification and label generation for radiology reports, which are then used to measure text similarity. When tested on a MIMIC dataset, this approach significantly outperformed traditional NLP metrics like ROUGE and BLEU, aligning more closely with clinical ground truth. This demonstrates the potential of LLMs for semantic analysis in highly specialized medical domains. It also entails the use of a human-in-the-loop for the post processing, something that our project is replacing with supervised fine-tuning on instruction tuned LLMs.

Elfes (2024) explored the possibilities of mapping text embeddings with news narratives using LLMs, employing structural linguistic techniques such as Greimas' Actantial Theory in order to represent news articles as narratives directed by six functional roles. The roles are generalizable and extracted with the help of LLMs, which are then integrated into a narrative-structure text embedding which was able to capture the semantic context and structure of text. This pipeline was then applied on news articles regarding the Israel-Palestine conflict and was able to distinguish articles that dealt with the same topics. The structural mapping of text embeddings with predefined narratives can be explored and extended onto the medical field as well, considering that we have a preset class of 14 disease labels. The functional roles mentioned in this paper can also be translated to medical reports, by breaking down chunks of texts into key functional components.

An approach to text summarization of clinical trial descriptions by Gulden et al. (2019) experimented with several text summarization algorithms on a corpus consisting of trial summaries and clinical descriptions. For evaluation, standard ROUGE metrics along with four independent reviewers assessed the content-completeness and inferred that extractive summarization was also helpful in preserving meaning of trial synopses. The hybrid approach to extract certain sections of the report first, and then proceed to fine tune the LLM based on these summarization to generate labels can be considered from this paper, and would include an additional insight into how the LLM decided to predict these labels.

Synthetic data offers promising applications in healthcare, from informing policy decisions to enhancing machine learning models for disease detection. Giuffrè and Shung (2023) highlight its potential in creating digital twins for hospital operations and patient treatment optimization. However, they also caution about challenges such as bias amplification and interpretability issues. The authors stress the importance of maintaining ethical standards and protecting patient privacy when utilizing synthetic data in healthcare applications. This section would be applicable during the instruction tuning phase - where synthetic data can be used to train the models, especially given the sensitive nature with respect to real life clinical trials. However, as the paper mentioned, the issues with bias amplification would lead to inherent biases and potential hallucinations which would hinder the performance of the model and also put people at risk.

Goel et al. (2023) outlines a two-step process for creating labels using annotated medical text. Combining the generative segments of the LLM along with the expertise of human raters, Goel et al. (2023) plans to feed the LLM with clinical reports and condition it to generate base labels, which are then inspected by a human professional to refine the corresponding generations. The main task of this approach was to determine the difference in time between the conventional approach i.e two rounds of human annotation to their method. This paper demonstrated a *yaml* format for the output generations which leads to efficient extraction of labels, and their setup of few shot on top of prompt engineering is something that will be incorporated in this project.

4 Dataset and Evaluation

4.1 Dataset

We utilize the CheXpert Plus dataset, a large public dataset of de-identified chest X-ray results. This dataset, developed by Chambon et al. (2024), provides over 200,000 text-based chest radiographic reports, making it suitable for training and evaluating our large language model on medical text analysis tasks. The CheXpert Plus dataset includes ground truth label annotations for each of the radiology reports obtained using the CheXbert labeler (Smit et al., 2020). Each report has annotations for 14 different chest pathologies, where each pathology is mapped to a positive mention, negative mention, uncertain mention, or no mention. The pathologies consist of the following 14 conditions: enlarged cardiomediastinum, cardiomegaly, lung opacity, lung lesion, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, support devices, no finding. Each report has a corresponding patient ID and study number. Figure 1 from Chambon et al. (2024) depicts the structure of the dataset mentioned above.

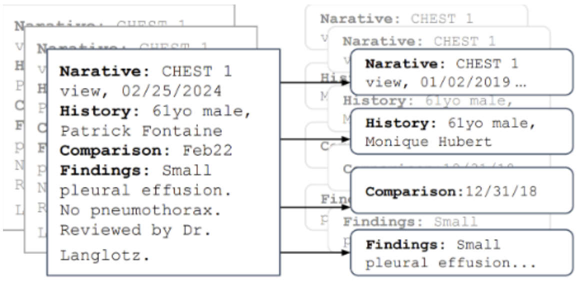

Narative: CHEST 1 view, 02/25/2024
History: 61yo male, Patrick Fontaine
Comparison: Feb22
Findings: Small pleural effusion. No pneumothorax. Reviewed by Dr. Langlotz.
Narative: CHEST 1 view, 01/02/2019 ...
History: 61yo male, Monique Hubert
Comparison:12/31/18
Findings: Small pleural effusion...

Figure 1: Dataset Overview

This dataset was filtered to remove radiology reports that do not contain a "findings" section. Since the dataset includes multiple images per radiology report, the dataset was filtered to remove duplicate samples containing the same patient ID and study number. Some patients also had identical reports across different studies, so we reduced the dataset to only include one report per patient. We also simplify the CheXbert multi-label classification using the method outlined in Chambon et al. (2024) so that each of the pathologies has two possible classes, positive and negative. Positive mentions and uncertain mentions are assigned to the positive class, since it better to run additional tests to diagnose a disease than completely rule it out. Negative mentions and missing mentions are assigned to the negative class.

We use an evaluation set of sizes 1,000 and 5,000 samples for our baseline model. While implementing the instruction tuned LLM, we perform a split of the entire dataset, with 70% of the dataset used for supervised fine-tuning and the remainder for testing.

4.2 Evaluation

To evaluate the performance of our model, we employ the exact match metric over the 14 class labels, since we primarily focus on label generation. The exact match metric measure success only when all of the 14 classes are predicted accurately. This is a necessary metric to evaluate on, as every single misclassification could be potentially life-threatening. Non-exact match accuracy also results in inconclusive labeling - if the predicted labels simultaneously label a radiology as having "No Findings" and also having one of the 13 conditions, then the results are not interpretable. Additional per-label evaluation metrics of precision, recall, and F1 score are used to describe the distribution of the model predictions. These metrics are calculated for every single column and then averaged over the number of labels(14). Here, we prioritize recall over precision, since it is better for the LLM to predict a positive label if uncertain, rather than predict a negative label and risk the chance of the disease being overlooked.

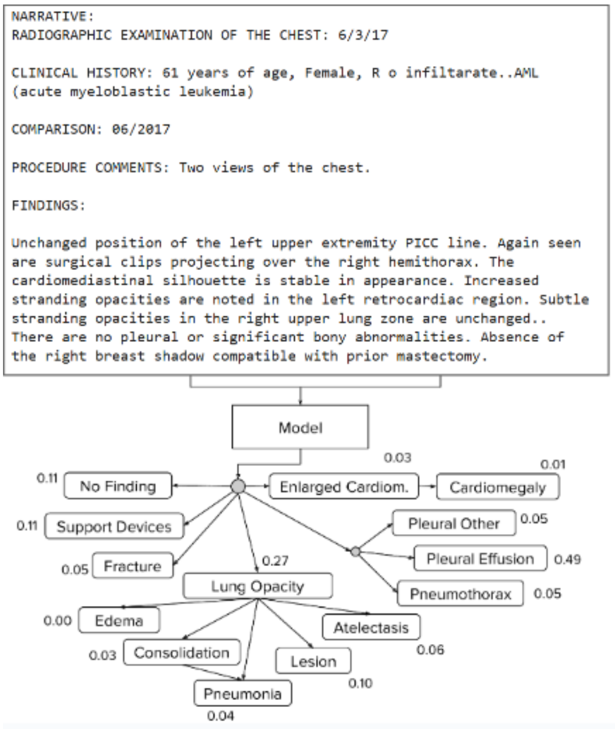
5 Methods

5.1 Baseline

For the baseline model, no training is done, so we use only use the evaluation set to test the model. The radiology reports are processed using an LLM encoder, specifically OpenAI's `text-embedding-3-small` model and the SBERT `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019), to generate feature vectors. These vectors are then compared with all the vectors from other reports using the cosine similarity score. The model takes the softmax of the cosine similarity scores, uses argmax to find the report that is most similar, and assigns the labels of the argmax report to the input report. This approach allows us to determine how closely a given report's semantics matches others in terms of content and inferred pathology.

5.2 Models

Adding on to the cosine similarity baseline, the next model proposed was using a base LLMs to generate these 14 vectors. The `Metal-Llama-3.1-8B` LLM was chosen for this task, as it is an open source model available to us. To begin with, the model was conditioned on a zero shot prompt and was instructed only to output the labels of the 14 labels in *yaml* format, which was proposed in Goel et al. (2023). This format allowed for easy extraction for evaluation of the labels using the metrics mentioned in section 4.2. The next step was to use a few shot setup, with one example from the dataset chosen for the same. This example was to show the model what segments of the report to focus on, and how some class labels were identified. However, it was decided to not use more than one example, since the primary source of text in this case is medical reports and hence the amount of data fed into the LLM should also be taken into account.



NARRATIVE:
RADIOGRAPHIC EXAMINATION OF THE CHEST: 6/3/17

CLINICAL HISTORY: 61 years of age, Female, R o infiltarate..AML (acute myeloblastic leukemia)

COMPARISON: 06/2017

PROCEDURE COMMENTS: Two views of the chest.

FINDINGS:

Unchanged position of the left upper extremity PICC line. Again seen are surgical clips projecting over the right hemithorax. The cardiomediastinal silhouette is stable in appearance. Increased stranding opacities are noted in the left retrocardiac region. Subtle stranding opacities in the right upper lung zone are unchanged.. There are no pleural or significant bony abnormalities. Absence of the right breast shadow compatible with prior mastectomy.

Figure 2: Model Pipeline

Based on the results from the base LLM, additional techniques were designed, with instruction tuning being the optimal choice. The *Meta-Llama-3.1-8B* LM was chosen again, with supervised fine tuning, proposed by Gunel et al. (2020) applied on the language model using a portion of the CheXpert Plus dataset as reference material.

6 Results

6.1 Baseline

In the baseline method, we used a 1,000 and 5,000 sample test set, and the accuracy was measured as the percentage of reports with correctly assigned labels over total samples. The accuracy is expected to be low because there are $2^{14} = 16384$ possible combinations of the 14 pathology labels, although practically, not all combinations are possible due to correlations between some of the labels (e.g. if No Findings is 1, all other pathologies would be 0). We also find the difference between the accuracies achieved when we find cosine similarities between full reports vs. using only the "impressions" field of these reports.

The results in Table 1 show the accuracy scores for the baseline model run on 1,000 and 5,000 samples and for an embedding of the entire report vs. only for the impression section of the report. The results in Table 2 shows the calculated accuracy score for the cosine similarity metric run on 1,000, 5,000, and 20,000 samples using the `SBERT` model embeddings. In the future, the `SBERT` embeddings will also be tested on the full reports. More samples were used when testing the `SBERT` embeddings due to prohibitive timing constraints when using the OpenAI API for the `text-embedding-3-small` embeddings. These results indicate room for improvement through proposed methods which aim at using both lexical and semantic similarity measures alongside traditional accuracy metrics.

Table 1: Baseline Accuracy Using `text-embedding-3-small` Vector Embeddings

Sample Count	Full Report	Impression Only
1000 Samples	0.048	0.068
5000 Samples	0.074	0.0664

Table 2: Baseline Accuracy Using `SBERT` Vector Embeddings

Sample Count	Impression Only
1000	0.057
5000	0.086
20000	0.0945

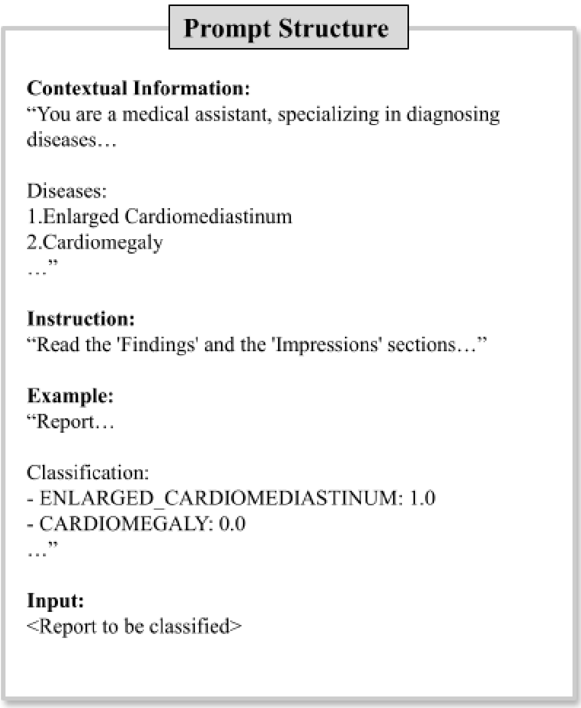
6.2 Base Llama 3

For the base Llama 3 model, a test set of 1,000 reports and impressions were sampled from the dataset. The exact match metric was chosen for this task as well, which measures success only when all the 14 disease labels are predicted accurately. Zero-shot and one-shot prompting were used on the model. For one-shot prompting, a single sample with a report and its corresponding labels was randomly selected from the dataset. The results of the model conditioned on the zero-shot and one-shot prompts respectively, are depicted in Table 3.

Table 3: Llama 3-8B Exact Match Accuracy

Prompt Style	Exact Match Accuracy
Zero-Shot	0.166
One-Shot	0.221

From these accuracies, we can see that there is a



Prompt Structure

Contextual Information:
“You are a medical assistant, specializing in diagnosing diseases…

Diseases:
1.Enlarged Cardiomediastinum
2.Cardiomegaly
…”

Instruction:
“Read the 'Findings' and the 'Impressions' sections…”

Example:
“Report…

Classification:
- ENLARGED_CARDIOMEDIASTINUM: 1.0
- CARDIOMEGALY: 0.0
…”

Input:
<Report to be classified>

Figure 3: One-Shot Prompt Structure

significant improvement from using just the cosine similarities of embeddings, which is to be expected. It also can be seen that the model conditioned on the one-shot is able to significantly outperform the zero-shot setup, which shows the effectiveness of the prompt setup to classify the 14 different disease labels.

Table 4: Llama 3-8B Per-Label Statistics

Prompt	Accuracy	Precision	Recall	F1
Zero-Shot	0.860	0.641	0.834	0.363
One-Shot	0.884	0.736	0.811	0.386

Table 4 shows the per-label statistics for the results of zero-shot and one-shot prompting. As expected, the per-label accuracy is much higher than the exact match accuracy, since exact match accuracy requires that all 14 labels match. For both zero-shot and one-shot, the recall is higher than precision, indicating that the model predicts more false positives than false negatives. This could be due to how the radiology reports are written. Radiology reports may include mentions of potential or suspected abnormalities that aren’t actually clinically significant, which can lead to predictions for conditions that don’t exist. There is an improvement in precision, i.e. a decrease in false positives, for one-shot compared to zero-shot, but no improvement in recall.

A possible reason for this is an imbalance in the dataset. In this radiology dataset, like in many medical datasets, negative results are more common than positive results for the 13 diagnoses. Thus, a randomly sampled few-shot setup could help improve the model’s prediction of true negatives, but have little effect on incorrect predictions of true positives.

6.3 Instruction Tuned Llama 3

From the results of the Base LLM, we can infer that there is a substantial increase in accuracy over using just the cosine similarity of the report embeddings. However, this accuracy is not optimal and can be fine tuned further to yield a better model. To improve the base LLM, we employed `Supervised Fine Tuning (SFT)` on the Llama model to adapt it to the task of medical report classification. Leveraging the *LoRA* framework, the goal was to efficiently update the parameters of the models while limiting the computational cost incurred. The parameters for *SFT* are explained below.

The process was initialized using *FastLanguageModel*, further augmented with LoRA adapters to target key modules such as the projection layers of the self-attention mechanism *(q_proj, k_proj, v_proj)*. Given the size of our dataset and the limited resources available, *lora_rank* and *lora_alpha* was not initialized to be too high, and was both set at 16. In addition, the gradient checkpoint technique was also employed, in order to support longer context sequences and allow for larger batch sizes while training, without significant VRAM overhead.

Training was then conducted using the *SFT-Trainer* module, and it was run for 60 steps with a linear learning rate scheduler of $2x10^{-4}$. Gradient accumulation was done across every 4 steps to store the weights. Mixed precision training through *fp16 and bf16* was applied to ensure accelerated computations, with `AdamW` being the optimizer of choice. These parameters were used for training, and the model reported a loss of ~0.43 averaged over 60 epochs. The model was then evaluated using the same prompts the base LLM was tested on, and the results are shown in tables 5 and 6.

From the exact match scores in table 5, we can observe that there is a substantial improvement in accuracy, both in the zero shot and one shot conditioned prompts. This can be attributed to the fact that the language model is now trained on

Table 5: Llama Instruct 3-8B Per-Label Statistics

Prompt	Exact Match Accuracy
Zero-Shot	0.291
One-Shot	0.305

the domain-specific data and as a result, is able to achieve higher scores in the exact match metric. These scores are also an indication that domain-adaptive knowledge drastically improves the reasoning capability of language models in that particular domain.

Table 6: Llama Instruct 3-8B Per-Label Statistics

Prompt	Accuracy	Precision	Recall	F1
Zero-Shot	0.906	0.830	0.727	0.775
One-Shot	0.917	0.834	0.746	0.791

The label accuracy metrics are depicted in table 6. From this, we can observe that `supervised fine tuning` has enhanced the ability to classify medical reports accurately, as there is a significant increase in all metrics. However, the recall metric has regressed from the base model. This could be due to how the prompt was structured, where uncertain cases or no mentions of class labels was assigned to the negative class. This would also explain the increase in precision from the previous attempt, jumping from *0.64* to *0.83*. Another factor, as previously mentioned in the 6.2 section, could be the class imbalance in the radiology dataset. As a result of these improvements, it can be seen that the F1-metric has almost doubled, from *0.36* to *0.775*, for zero shot and *0.386* to *0.791* for the one shot conditioned prompts.

A comprehensive comparison of the exact match metric of the different methods experimented on is illustrated in Figure 4.

7 Future Work

7.1 Improvements

A possible improvement that can be made to the methods in this project is to have more extensive prompt engineering. For example, the prompt could be modified to tell the model what to do in the case of uncertainty, i.e. to lean towards either predicting the positive or the negative class.

Another improvement could be leveraging self-instruct Wang et al. (2023) and reinforcement-based finetuning methods in order to reduce variability. The model can also be prompted to regenerate the prediction if it does not make logical sense (i.e. a result of *No Finding* and another class).

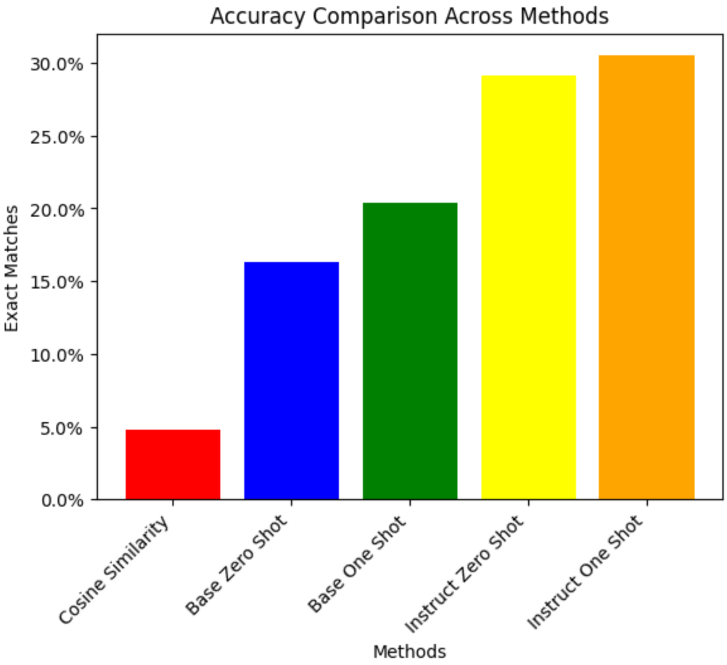


Figure 4: Method Comparison over 1000 samples

7.2 Extensions

The exploration of the hypothesis focuses on different methods for classifying medical conditions by inferring from chest radiographs. While this has demonstrated encouraging results, there is significant potential for extending the applications of this project, which will be discussed in this section.

Since this project only deals with medical reports related to chest radiographs, the next logical step would be expanding the dataset to incorporate diseases that affect other different body parts as well, which can be found by scans or tests other than X-rays. Broadening disease coverage also results in a more generalized disease classification language model.

It can be further extended to consider other categories of medical or clinical texts. Radiology reports like MRI and X-ray findings follow specific structures that differ from more general clinical notes. These free-text clinical notes could also benefit from LLM acceleration of text classification.

Adding on to the above point, to achieve this, a copious amount of data will be required. This data will be used for various purposes, ranging from training to supervised fine-tuning to evaluation. An extension of the study conducted by Gulden et al. (2019) on techniques to construct synthetic data of benchmark quality will accelerate the process of creating generalized LMs for disease classification.

Another avenue for future exploration would be to diversify the types of instructions we want the language model to answer. Instead of just classifying the presence/absence of a particular disease, an triage assessment to explain the severity of the disease can also be included in the prompt. Other ideas for instructions can be multimodal instructions, taking in information from previous reports, the patients history to make a decision, and also as a tool concatenated to scanning devices which can aid in highlighting key abnormalities present in the body.

References

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

Jan Elfes. 2024. Mapping news narratives using llms and narrative-structured text embeddings. *arXiv preprint arXiv:2409.06540*.

MAuro Giuffrè and Dennis L Shung. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine*, 6(1):159.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. Llms accelerate annotation for medical information extraction. *Preprint*, arXiv:2312.02296.

Christian Gulden, Melanie Kirchner, Christina Schüttler, Marc Hinderer, Marvin Kampf, Hans-Ulrich Prokosch, and Dennis Toddenroth. 2019. Extractive summarization of clinical trial descriptions. *International journal of medical informatics, 129:114–121*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pretrained language model fine-tuning.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Shaochen Xu, Yiwen Jiang, Yijun Ren, and Xiaoqian Jiang. 2024. Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis. *arXiv preprint arXiv:2402.11398*.