

CXR Fact Encoder: Combining Large Language Models with Medical Knowledge for Enhanced Radiological Text Representation

Anonymous ACL submission

Abstract

Recent advancements in representation learning, although promising, often confront challenges in specialized domains like medicine. In particular, the acquisition of expert annotations for medical texts and images is notably burdensome due to the limited availability and time constraints of medical professionals. Recognizing this, Large Language Models (LLMs) offer a promising avenue to automatically extract annotations from radiology reports at scale. In this work, we exploit the potential of pairing LLMs with domain-specific knowledge, thus reducing the dependency on time-intensive human expert annotations for improved medical text representation. Specifically, we introduce a two-stage system for the extraction and encoding of facts from radiology reports using LLMs such as ChatGPT and T5, in tandem with specialized medical knowledge sources. As a cornerstone of this system, we present CXR Fact Encoder—a BERT-based model fine-tuned for the enhanced representation of chest X-ray radiology reports. Additionally, we illustrate the applicability of our method by introducing *CXR Fact Encoder Score*, a novel evaluation metric crafted specifically for radiology text generation, drawing from all the elements of our two-stage system. Our evaluations show the proposed system outperforms multiple baseline methods in tasks like sentence ranking, natural language inference, and label extraction from radiology reports. We make our model weights, data, and code publicly available.

1 Introduction

In the context of medical image analysis, radiology reports constitute a rich source of unstructured information. Such free-text radiology reports are written by radiologists as part of their regular practice and are typically comprised of sections such as *comparison*, *indication*, *findings*, and *impression*. Figure 1 shows an illustrative example of



Comparison: Chest radiographs XXXX.
Indication: XXXX-year-old male, chest pain.
Findings: The cardiomeastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Stable calcified granuloma within the right upper lung. No acute bone abnormality.
Impression: No acute cardiopulmonary process.

Figure 1: Example image and report from the IU X-ray dataset (Demner-Fushman et al., 2015)

such reports in the context of Chest X-ray (CXR) images.

Radiology reports can be utilized in different manners. One use case is label extraction to provide structured supervision for medical image tasks, such as abnormality classification or detection (Irvin et al., 2019; Smit et al., 2020; Jain et al., 2021b; Bustos et al., 2019; Syeda-Mahmood et al., 2020; Wu et al., 2021; Jain et al., 2021a). Other use cases include radiology report generation (Messina et al., 2022; Miura et al., 2021; Delbrouck et al., 2022; Tanida et al., 2023) and summarization (Chen et al., 2023b; Ma et al., 2023). Another recent trend is the development of multimodal models that can jointly understand medical images and text using different techniques, such as image and text masking and contrastive learning (Wang et al., 2022; Boecking et al., 2022; Bannur et al., 2023; Moon et al., 2022; Chen et al., 2022).

For all these tasks, a key step is the correct understanding of the factual information contained in the report. In particular, the *findings* and *impression* sections of a report can be viewed as a collection of facts about the imaging exam. Facts may include observations (of abnormalities, diseases, devices, etc.), an interpretation or inference from one or more observations, references to some anatomical location, discussions of the level of severity or degree of confidence, comparisons with respect to a previous study, etc. For example, in Figure 1, one fact is that there is *no acute bone abnormal-*

ity (a normal observation), and another fact is that there is *stable calcified granuloma within the right upper lung* (an abnormal observation in a specific anatomical location).

The lack of suitable methods for fact extraction and encoding for medical reports motivates us to develop a new method to tackle this problem. Specifically, our proposed method can extract medical facts, encoding them into a high-quality latent representation that captures clinical details while accounting for variations in radiology report free-text. Our approach is also inspired by the capabilities of Large Language Models (LLMs) like GPT-3.5 and GPT-4—often referred to as versions of ChatGPT—which have demonstrated exceptional medical performance (Katz et al., 2023; Liu et al., 2023b; Adams et al., 2023). We also leverage insights from expert-annotated datasets, including Chest ImaGenome (Wu et al., 2021), RadGraph (Jain et al., 2021a), MedNLI (Romanov and Shivade, 2018), and RadNLI (Miura et al., 2021).

Paper contributions. In light of these motivations, our work presents the following contributions:

- A fact extractor: a novel and simple approach to extracting facts from Chest X-ray radiology reports by leveraging LLMs. We use ChatGPT and a fine-tuned version of T5 (Raffel et al., 2020) in order to capture relevant information from reports, without requiring annotations from radiologists.
- A fact encoder: *CXR Fact Encoder* for CXR reports. The model is based on the BERT architecture and shares the same tokenizer and initial weights as CXR-BERT-specialized (Boecking et al., 2022), but is further fine-tuned with a multi-task supervisory approach that leverages domain expertise from radiologists as well as ChatGPT and T5 generated annotations. As a result, CXR Fact Encoder exhibits significant advancements in fact comprehension, as demonstrated by improved sentence ranking and natural language inference capabilities. Moreover, the entire system (fact extraction + encoding) can be used for label extraction from reports, outperforming several baselines.
- A new evaluation metric for radiology text generation, that we name *CXR Fact Encoder Score*, that measures the factual correctness

of a generated text with respect to a ground-truth text, by extracting and comparing the embeddings of the facts in each one. This is one of the many possible applications of our two-stage system.

We release the weights of CXR Fact Encoder, the weights of the fine-tuned version of T5 for fact extraction, as well as data and code necessary to replicate the results. We also release *CXR Fact Encoder Score* as a Python library for ease of use by the research community.

Paper organization. The remainder of the paper is structured as follows: Section 2 explores related work, emphasizing BERT-based radiology text representation, label extraction, factual correctness in radiology text generation, and LLMs. Then Sections 3 and 4 present the two stages of our proposed system, namely, fact extraction and fact encoding, respectively. Section 5 describes the datasets used in our experiments, including details about our annotation strategy. Our experimental evaluation is captured in Section 6, where we present various tasks, emphasizing the efficacy of our approach. We conclude in Section 7 with key insights and contributions, while Section 8 acknowledges limitations and suggests future research avenues.

2 Related Work

BERT for Radiology Text Representation. In recent years, BERT (Devlin et al., 2019) has revolutionized various domains of natural language processing (NLP), offering remarkable improvements in text representation. Consequently, subsequent works have developed new variants of BERT for different text-related applications. Some examples in the medical domain are BioClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2020), BioLinkBERT (Yasunaga et al., 2022), CXR-BERT (Boecking et al., 2022) and BioViL-T (Bannur et al., 2023). Like these works, we follow the common practice of making BERT the basis of our model. However, our work differs in the fact that we follow a different training protocol that takes advantage of LLMs like ChatGPT to generate supervision at large scale, in addition to supervision obtained from datasets annotated by domain experts.

Label extraction from Radiology Reports. Our work is also related to the problem of extracting information, usually in the form of labels, from free-text radiology reports. A well-known exam-

ple in the literature is the CheXpert labeler (Smit et al., 2020), which uses a rule-based system to infer the presence or absence of 13 observations (plus the label "No findings"). CheXbert (Smit et al., 2020) and VisualCheXbert (Jain et al., 2021b) are subsequent versions that follow the same labeling standard of CheXpert but are based on BERT. The Chest ImaGenome dataset (Wu et al., 2021) is another example that made use of a rule-based NLP system to label reports in order to build scene graphs for the corresponding frontal images in the MIMIC-CXR dataset (Johnson et al., 2019a). RadGraph (Jain et al., 2021a) proposed a labeling standard of entities and relations for radiology reports, and trained a variant of BERT, DyGIE++ (Wadden et al., 2019), for entity and relation extraction on examples annotated by radiologists. PadChest (Bustos et al., 2019) followed a similar approach, by labeling Spanish reports with a LSTM that was previously trained on examples annotated by physicians. Our work contributes in this domain by proposing a different method for information extraction, by combining the powerful representation capabilities BERT with the remarkable natural language skills of ChatGPT and T5, in order to extract and encode facts from reports.

Factual Correctness in Radiology Text Generation. One important area of application motivating this work is the evaluation of factual correctness in systems that generate radiological text. Recent works have stressed the importance of improving and optimizing the correctness of the facts generated by a system in applications such as report generation (Miura et al., 2021; Delbrouck et al., 2022) and summarization (Zhang et al., 2020b). Likewise, Yu et al. (2022) conducted a study on metrics to evaluate progress in automatic CXR report generation, and concluded that the best ones were all based on BERT. Thus, a direct application of our work is the use of CXR Fact Encoder as a learned metric of medical factual correctness, by extracting and comparing facts in a latent space.

LLMs in Medicine. Our work falls under the category of applications of LLMs to the medical domain. Specifically, in this work we make use of ChatGPT versions GPT-3.5 and GPT-4 through OpenAI’s API¹. Previous works have successfully applied ChatGPT to medical tasks. Liu et al. (2023b) used ChatGPT to generate short sentences with plausible symptoms of medical conditions for

interpretable zero-shot medical image diagnosis. Adams et al. (2023) used GPT-4 to transform free-text radiology reports into structured templates, with remarkable results. GPT-4 is also known for having passed the bar exam (Katz et al., 2023).

Knowledge Distillation from LLMs. Our approach can be also viewed as a form of LLM knowledge distillation, where a LLM ("teacher") is queried to generate annotations for training a more compact model ("student"). Shi et al. (2023) illustrated this idea by using ChatGPT to extract knowledge graphs from text to train a smaller model for text classification. Similarly, Gu et al. (2023) applied this concept in the biomedical field, distilling knowledge from GPT-3.5 for adverse drug event extraction, with student models like PubMedBERT and BioGPT.

3 Fact Extraction

Figure 2 outlines our method for extracting facts from radiology reports, with an example from the MIMIC-CXR dataset (Johnson et al., 2019b). Initially, we use regular expressions and simple rules to pinpoint relevant radiological sections in MIMIC-CXR reports, mainly *Findings* and *Impression*, but we also handle alternate headings. These sections are then divided into sentences. For simplicity, we use the `sent_tokenize` function from the NLTK library², resulting in 677,694 unique sentences after processing the entire dataset. Finally, we retrieve facts from each sentence. The rationale for this is that radiologists occasionally compose intricate sentences that encapsulate multiple observations. As an example, Figure 2 demonstrates a sentence conveying three distinct facts. Given the recent success of Large Language Models, an effective strategy to achieve this extraction is by directing an LLM, like ChatGPT, using a custom prompt. The precise prompt and an example are provided in Figure 11 in the Appendix.

In principle, this entire stage could be accomplished by LLMs. However, we faced a challenge due to the high costs associated with using pay-per-use APIs for LLMs, which can escalate significantly for large text annotation tasks. A solution is to annotate a strategic subset of sentences with a costly LLM and then distill the knowledge captured by these annotations into a more affordable sequence-to-sequence model, such as T5, via fine-tuning. As a precedent, this strategy is similar

¹<https://platform.openai.com/>

²<https://www.nltk.org/>

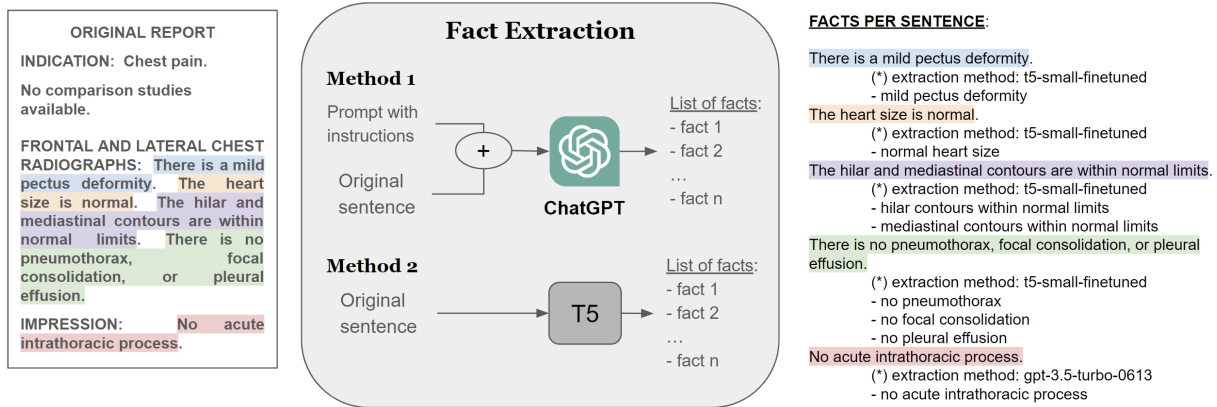


Figure 2: Fact extraction procedure for radiology reports.

to Yang et al.’s approach (2023) where they fine-tuned T5 to condense GPT-3’s verbose descriptions in LLM-assisted image classification. In our case, we annotated 14,999 sentences with GPT-4-0613, 69,936 with GPT-3.5-turbo-0613, and used T5-small for the remaining 592,759 sentences after its fine-tuning. This process resulted in 1,323,687 facts, including duplicates, and 583,202 unique facts post-duplicate removal.

4 CXR Fact Encoder

After we extract facts, we encode them by representing them as vectors in a latent space via a text embedding model, which we refer to as *CXR Fact Encoder*. In our experiments we rely on CXR-BERT (Boecking et al., 2022) to implement our fact encoder. Specifically, we use the CXR-BERT-specialized variant available on the Huggingface hub³. CXR-BERT is a BERT-based text encoder with a domain-specific tokenizer for CXR reports. It was trained with three phases of pretraining that include masked language modeling, radiology section matching, regularisation, and text augmentations. CXR-BERT-specialized is a version that is further fine-tuned via a multimodal contrastive learning framework that matches CXR images and reports, similar to the CLIP framework (Radford et al., 2021), so that the latent representation of the [CLS] token is used to align text/image embeddings.

Building on top of CXR-BERT-specialized, we explore 6 different approaches to enhancing the latent representation of radiological sentences: triplet loss for sentence ranking (T), natural language inference (NLI), quadruplet loss to enforce a sepa-

³<https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-specialized>

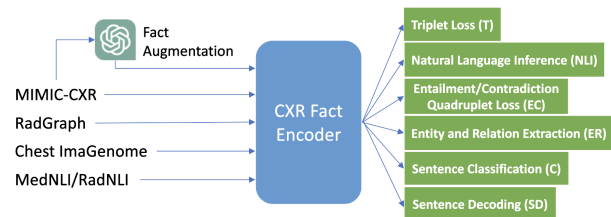


Figure 3: CXR Fact Encoder model.

ration between entailment and contradiction pairs (EC), entity and relation extraction (ER), sentence classification tasks (C), and sentence decoding (SD). Figure 3 presents a high level summary of the different tasks and datasets used to fine-tune the model. Details for each task’s implementation can be found in Section A.1 of the Appendix.

Thus, by combining the two stages, the whole system can accurately extract and encode facts from CXR reports, thus providing a rich and convenient representation of the factual information for downstream applications.

5 Datasets

The primary dataset used in our experiments is MIMIC-CXR (Johnson et al., 2019b), which we already alluded to in the previous sections. This dataset comes with 227,827 radiology reports associated with 377,110 chest X-ray images. In this paper we only carry out experiments using the reports, leaving the use of images and multi-modality for future work.

In addition, we utilize annotations from the Chest ImaGenome (Wu et al., 2021) dataset. Chest ImaGenome was created to offer a relatively broad label set that links multiple observations to anatomical image locations. For every one of the 242,072

frontal view images in MIMIC-CXR, Chest Im-
aGenome gives a scene graph connecting report ob-
servations to image anatomical locations via bound-
ing boxes. This is achieved through a combined
rule-based NLP and atlas-based bounding box de-
tection process, backed by a CXR lexicon and on-
tology crafted with radiologists’ help. We leverage
this dataset for two main reasons: 1. The scene
graphs pinpoint the exact report sentence where
each observation and location are derived, facilitat-
ing the creation of a binary multi-label classifi-
cation task for a text encoder to predict sentence-
based observations and locations. 2. Chest Im-
aGenome introduces a radiologist-informed an-
notation standard, covering 74 observation types
(grouped in categories such as *anatomical finding*,
disease, and *texture*) and 38 anatomical locations
(e.g., *right upper lung zone* and *spine*).

Similarly, RadGraph (Jain et al., 2021a) provides
graph-based annotations for radiology reports. A
subset of 500 MIMIC-CXR reports were manually
annotated by board-certified radiologists using a
specific entity and relation schema. These anno-
tations trained a DyGIE++ model (Wadden et al.,
2019) for entity and relation extraction, which then
automatically annotated the rest of the MIMIC-
CXR reports. The dataset also features a test set
comprising 100 MIMIC-CXR and CheXpert re-
ports, each annotated by two board-certified radi-
ologists, and 500 CheXpert reports annotated by
DyGIE++. RadGraph is incorporated into our ex-
periments due to its rich domain-specific annota-
tions.

As we conduct experiments on NLI, we also
leverage the datasets MedNLI (Romanov and Shiv-
ade, 2018), RadNLI (Miura et al., 2021) and MS-
CXR-T (Bannur et al., 2023), which we describe
in more detail in Section 5.3. We also use the IU
X-ray dataset (Demner-Fushman et al., 2015) for
sentence ranking evaluation, as we will explain in
Section 6.

5.1 Extracting and annotating facts

We first run the fact extraction procedure presented
in Section 3. After that, we enhance these anno-
tations in several ways. We employ ChatGPT to
generate paraphrases of the facts, an example of
which is in Figure 12 in the Appendix. Each fact is
further annotated with a JSON metadata object, en-
compassing fields like "anatomical location", "de-
tailed observation", "short observation", "category",
"health status" and "comparison status". The re-

spective prompt for this is in Figure 14. To refine
the "comparison status" field, we utilize another
prompt displayed in Figure 15. Furthermore, we
prompt ChatGPT to label in line with the Chest Im-
aGenome dataset’s annotation standards, as demon-
strated in Figures 16 and 17. Notably, when adding
metadata and Chest ImGenome labels to facts, we
adopt the approach detailed in Section 3: we se-
lectively use ChatGPT for a subset and then train
T5 for the remaining annotations. This method lets
us expand Chest ImGenome annotations to more
sentences than originally included in the dataset.

5.2 Triplet Sampling Heuristics

CXR Fact Encoder is trained to generate sentence
embeddings that cluster semantically similar sen-
tences in the embedding space through a triplet
ranking task with binary cross-entropy loss. This
approach uses a dataset of triplets, each one with an
anchor, a positive sample (akin to the anchor), and a
negative one. The difference in similarities is com-
puted as $\Delta\text{sim}(a, p, n) = \text{sim}(a, p) - \text{sim}(a, n)$
from their embeddings’ dot product. By minimiz-
ing the binary cross-entropy loss, the encoder en-
sures closely related sentences are nearer and unre-
lated ones are more distant in the embedding space.

To define our triplet sampling heuristics, we use
the notation $E(x)$ for the embedding of sentence x ,
 $\cos(E(x), E(y))$ for the cosine similarity between
embeddings of x and y , $\text{lev}(x, y)$ for the leven-
shtein string distance between them, and $\text{levsim}(x,$
 $y) = 1 - \text{lev}(x, y) / \max(\text{len}(x), \text{len}(y))$. $c(x)$ indi-
cates the cluster id for sentence x after running a
clustering algorithm like K-Means on the sentence
embeddings. With this, we sample triplets based
on these heuristics:

Rule 1: Rank paraphrases very high.
 $\Delta\text{sim}(a, p, n) > 0$ if p is a paraphrase of a gen-
erated by ChatGPT and n is any other sentence (un-
less $\cos(E(a), E(p)) < \cos(E(a), E(n))$ and $\text{lev}(a,$
 $p) > \text{lev}(a, n)$).

**Rule 2: Sample triplets according to the con-
sensus of E and lev, while anchor and positive
share the same health status.** $\Delta\text{sim}(a, p, n) >$
 0 if $\text{HS}(a) = \text{HS}(p)$, $c(p) = c(a)$, $c(p) \neq c(n)$,
 $\cos(E(a), E(p)) > \cos(E(a), E(n)) + \text{margin}_{\cos}$ and
 $\text{levsim}(a, p) > \text{levsim}(a, n) + \text{margin}_{\text{lev}}$.

**Rule 3: Short observation, detailed obser-
vation and the original fact (and their para-
phrases) should be close to each other.** Given
a fact f , $\Delta\text{sim}(a, p, n) > 0$ if a and $p \in S(f)$, n
 $\notin S(f)$ and $c(a) \neq c(n)$ (unless $\cos(E(a), E(p)) <$

$\cos(E(a), E(n))$ and $\text{lev}(a, p) > \text{lev}(a, n)$). Here, $S(f)$ stands for the union of f , its detailed observation, its short observation and all the paraphrases (if any) generated for all of them with ChatGPT.

Rule 4: Sample triplets according to Chest ImaGenome labels. $\Delta\text{sim}(a, p, n) > 0$ if $\text{CIGL}(a) \cap \text{CIGL}(p) \neq \emptyset$, $\text{CIGL}(a) \cap \text{CIGL}(n) = \emptyset$, $\text{CIGL}(p) \cap \text{CIGL}(n) = \emptyset$, and if $(\cos(E(a), E(p)) > \cos(E(a), E(n)))$ AND $\text{levsim}(a, p) > \text{levsim}(a, n)$. Here, $\text{CIGL}(x)$ stands for the set of Chest ImaGenome labels of the sentence x .

Rule 5: Rank triplets according to the overlap of entities and relations from RadGraph. $\Delta\text{sim}(a, p, n) > 0$ if $c(a) = c(p)$, $c(a) \neq c(n)$, and $J(\text{RG}(a), \text{RG}(p)) > J(\text{RG}(a), \text{RG}(n)) + \text{margin}_{\text{RG}}$. Here, $\text{RG}(x)$ stands for the set of RadGraph entities and relations for the sentence x , and J for Jaccard similarity.

Rule 6: Hard triplets generated by ChatGPT. $\Delta\text{sim}(a, p, n) > 0$ if (a, p, n) is a hard triplet generated by ChatGPT. Figure 18 shows the prompt used to generate these triplets along with an example.

For each rule, we create approximately 3 million training triplets, and 1,000 each for validation and testing. Rule 1 additionally involves generating paraphrases for anatomical location sentences, with the prompt displayed in Figure 13. Many of these rules utilize an auxiliary embedding for sentence clustering and cosine similarity. In our experiments, we choose BioViL-T (Bannur et al., 2023), an advanced version of CXR-BERT available on Huggingface⁴. This version retains the original architecture but offers enhanced comprehension of temporal text descriptions.

5.3 Natural Language Inference

Natural Language Inference (NLI) classifies the relationship between a premise and a hypothesis into "entailment", "neutral", or "contradiction". For instance, in a CXR report, a premise might state "There are no evident signs of pleural effusion", while a hypothesis says "There are evident signs of pleural effusion". Although structurally similar, they contradict each other, emphasizing the importance of nuanced comprehension in radiology reports. The goal of using NLI during training is to perfect sentence embeddings at detecting these subtle distinctions.

For training, all MedNLI splits (Romanov and Shivade, 2018) are used, amounting to 14,049 an-

notated medical sentence pairs. Radiology-specific datasets include RadNLI (Miura et al., 2021) with 960 pairs and MS-CXR-T (Bannur et al., 2023), an evaluation set with 361 pairs emphasizing condition evolution over time. Given the limited NLI samples from CXR reports, the RadNLI development set (480 pairs) is used for training, and the rest is left for evaluation. To enrich the training dataset, we use GPT-4 to obtain 147,509 new pairs using four distinct prompts (see Figures 19, 20, 21, 22 in the Appendix), resulting in a total of 162,036 pairs categorized as 26,442 entailment, 39,817 neutral, and 95,777 contradiction pairs.

6 Experimental Results

In the majority of our experiments, we assess various versions of CXR Fact Encoder. Each version is trained on two or more of the tasks listed in Figure 3. For triplet loss, we employ the loss function and dataset described in Section 5.2. The classification tasks include category (5 classes), health status (4 classes), comparison status (15 classes), Chest ImaGenome observations (74 classes) and anatomical locations (38 classes). For RadGraph entity and relation extraction we augment CXR Fact Encoder with SpERT (Ebarts and Ulges, 2020). For sentence decoding, we attach a lightweight transformer decoder to the model. We refer the reader to Section A.1 in the Appendix for a more detailed description of each task.

Triplet and Sentence Ranking. We evaluate CXR Fact Encoder and multiple baselines from the literature on triplet ranking accuracy. We also report AUC on a sentence ranking evaluation with 8617 sentences from IU X-ray reports. In this evaluation, given two sentence x and y , we heuristically say that y is relevant for x if $J(\text{RG}(x), \text{RG}(y)) \geq 0.4$ or $(J(\text{RG}(x), \text{RG}(y)) \geq 0.2$ and $(\text{CXP}(x) = \text{CXP}(y)$ or $\text{CXB}(x) = \text{CXB}(y)))$. Here J stands for Jaccard, RG for RadGraph entities and relations, CXP for CheXpert labels and CXB for CheXbert labels.

Table 1: Triplet and sentence ranking results.

ID	Text Model	Triplets Test Set (1000 samples per rule)						IU X-ray AUC	
		R1 (obs)	R1 (anat)	R2	R3	R4	R5		R6
1	BioLinkBERT (Yasunaga et al., 2022)	0.753	0.725	0.786	0.756	0.644	0.774	0.520	0.862
2	PubMedBERT (Gu et al., 2020)	0.901	0.853	0.905	0.873	0.767	0.834	0.603	0.908
3	BioClinicalBERT (Absentzer et al., 2019)	0.922	0.864	0.935	0.912	0.834	0.948	0.601	0.924
4	CheXbert (Smit et al., 2020)	0.855	0.771	0.908	0.884	0.760	0.937	0.635	0.933
5	CXR-BERT-specialized (Bocking et al., 2022)	0.880	0.804	0.992	0.914	0.904	0.932	0.717	0.852
6	BioViL-T (Bannur et al., 2023)	0.910	0.851	1.000	0.938	1.000	0.944	0.765	0.866
7	CXR Fact Encoder (T)	0.968	0.955	0.925	0.964	0.798	0.952	0.946	0.914
8	CXR Fact Encoder (T+C)	0.967	0.945	0.967	0.982	0.926	0.988	0.937	0.944
9	CXR Fact Encoder (T+R)	0.962	0.946	0.917	0.961	0.798	0.954	0.927	0.904
10	CXR Fact Encoder (T+SD)	0.981	0.966	0.954	0.977	0.875	0.981	0.898	0.953
11	CXR Fact Encoder (T+EC)	0.963	0.952	0.942	0.969	0.797	0.964	0.942	0.807
12	CXR Fact Encoder (T+EC+NLI)	0.941	0.944	0.925	0.945	0.751	0.936	0.919	0.758
13	CXR Fact Encoder (T+C+EC+NLI+ER)	0.976	0.948	0.969	0.980	0.905	0.979	0.929	0.901
14	CXR Fact Encoder (T+C+EC+NLI+SD)	0.973	0.964	0.976	0.989	0.905	0.982	0.940	0.909

Table 1 presents the results. Notably, all different

⁴<https://huggingface.co/microsoft/BiomedVLP-BioViL-T>

versions of CXR Fact Encoder outperform all the baselines in triplet rules where ChatGPT is heavily involved, namely, paraphrases (R1, R3) and hard triplets (R6). BioViL-T achieves perfect scores in R2 and R4 but this is by design, as BioViL-T is used as auxiliary embedding in triplet sampling (see Section 5.2). Sentence decoding (SD) and classification (C) appear to be helpful auxiliary tasks since most of the best scores are achieved by variants that include them (rows 8, 10, 14).

NLI. Table 2 shows NLI results using cosine similarity between sentence vectors, following a similar evaluation protocol as in Bannur et al. (2023). Only entailment and contradiction pairs are considered, excluding RadNLI’s neutral pairs. Results are determined based on a similarity threshold. Notably, the use of the entailment/contradiction quadruplet loss (rows 11-14) is key for top performance, significantly outperforming all the baselines, whereas variants without EC (rows 7-10) show weaker separation.

Table 3 displays accuracy on the RadNLI test set, including RadNLI’s neutral pairs (280), along with entailment (102) and contradiction (98) pairs. In this setting, the NLI classification head of CXR Fact Encoder is applied (refer to Figure 9). CXR Fact Encoder fine-tuned solely for NLI scores 79.8, practically equal to PTUnifier’s 80.0 and just slightly behind DoT5 (82.1), which follows a sophisticated sequence-to-sequence approach based on T5. CXR Fact Encoder (T+C+EC+NLI+SD) closely follows with 78.1. To estimate an upper bound for how much NLI knowledge could be distilled from GPT-4, we test its performance using the prompt in Figure 21. GPT-4 achieves 82.3, which to the best of our knowledge would be considered SOTA, although only marginally better than the other methods. For further inspection, Figure 4 provides confusion matrices for both CXR Fact Encoder and GPT-4, highlighting good distinction between contradiction and entailment but some confusion with neutral pairs.

Label extraction. We evaluate our two-stage system (ChatGPT/T5 + CXR Fact Encoder) against three radiology report label extraction methods: CheXpert labeler (Irvin et al., 2019), CheXbert (Smit et al., 2020), and Chest ImaGenome (Wu et al., 2021). For Chest ImaGenome, we use the labels from the dataset’s scene graphs, as the original NLP algorithm is not publicly available. We created an evaluation protocol to measure factual correctness and completeness: for each MIMIC-

CXR test set report and label extraction method, labels are extracted, converted into a report using templates, and then evaluated against the original report using report generation metrics. For CheXpert labeler and CheXbert we employ the templates suggested by Pino et al. (2021), while Chest ImaGenome uses basic templates like “(no) {observation} in {anatomical location}”. CXR Fact Encoder employs a label extraction method based on K-Medoids clustering of fact and anatomical location embeddings, resulting in labels represented as pairs (*fact_cluster_id*, *anatomy_cluster_id*) or just *fact_cluster_id* if an anatomical location is not available for the fact. Reports are generated from these labels using representative sentences from our dataset. Further procedure details can be found in Section A.2 in the Appendix. Table 5 provides examples of template-based reports.

CXR Fact Encoder Score. As part of the evaluation, we introduce *CXR Fact Encoder Score*. Given a reference and generated report, we extract facts from each and represent them as embedding vectors, denoting the sets for the original and generated reports as O and G respectively. The cosine similarity matrix M of size $|O| \times |G|$ is formed, where $M_{i,j}$ represents the cosine similarity between the i^{th} vector of O and the j^{th} vector of G . Using a similarity threshold t , we compute precision (P), recall (R), and F1-score (F_1). A "soft" version of the metric calculates average similarities S_{row} , S_{col} , and S .

$$\begin{aligned}
 P &= \frac{\sum_j \mathbb{1}(\max_i M_{i,j} \geq t)}{|G|} & S_{col} &= \frac{\sum_j \max_i M_{i,j}}{|G|} \\
 R &= \frac{\sum_i \mathbb{1}(\max_j M_{i,j} \geq t)}{|O|} & S_{row} &= \frac{\sum_i \max_j M_{i,j}}{|O|} \\
 F_1 &= 2 \times \frac{P \times R}{P + R} & S &= \frac{S_{row} + S_{col}}{2}
 \end{aligned}$$

Label extraction results. Table 4 presents results of template-based report generation using various label extraction methods. We report results with the new *CXR Fact Encoder Score* and also include RadGraph metrics (Jaccard similarity, F1 score, Precision, Recall), CheXpert and CheXbert metrics (accuracy, F1 macro average). Notice that CXR Fact Encoder, CheXpert labeler, and CheXbert are applied in both label extraction and evaluation. In addition, we report BERTScore (Zhang et al., 2020a), BLEU (Papineni et al., 2002), CIDEr-D (Vedantam et al., 2015), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). We observe a consistent improvement across all metrics using CXR Fact Encoder as the number of clusters and labels increases. As an upper bound,

Table 2: Results on NLI as sentence similarity. Acc_{E+C}^* denotes an upper bound in the accuracy with an optimal similarity threshold tuned in the same evaluation data.

ID	Text Model	NLI mixed dataset (Used to adjust similarity threshold) Entailment pairs: 26,442 Contradiction pairs: 95,777				RadNLI test set Entailment pairs: 102 Contradiction pairs: 98				MS-CXR-T Entailment pairs: 141 Contradiction pairs: 220			
		Best threshold	Acc _E	Acc _C	Acc _{E+C}}	Acc _E	Acc _C	Acc _{E+C}}	Acc _{E+C}^*}	Acc _E	Acc _C	Acc _{E+C}}	Acc _{E+C}}
1	PubMedBERT (Gu et al., 2020)	0.833	100.0	0.0	50.0	100.0	0.0	50.0	63.8	100.0	0.0	50.0	56.5
2	BioLinkBERT-large (Yasunaga et al., 2022)	1.000	0.0	100.0	50.0	0.0	100.0	50.0	65.9	0.0	100.0	50.0	58.1
3	BioClinicalBERT (Alsentzer et al., 2019)	0.660	100.0	0.0	50.0	100.0	0.0	50.0	69.2	100.0	0.0	50.0	69.5
4	CheXbert (Smit et al., 2020)	0.598	85.6	34.5	60.1	91.2	85.7	88.4	90.4	100.0	1.8	50.9	63.3
5	CXR-BERT-specialized (Boecking et al., 2022)	0.713	74.0	49.2	61.6	59.8	89.8	74.8	82.3	100.0	12.2	56.1	77.5
6	BioVIL-T (Banmur et al., 2023)	0.735	73.0	52.0	62.5	58.8	93.9	76.3	77.9	100.0	10.0	55.0	87.8
7	CXR Fact Encoder (T)	0.713	79.1	58.3	68.7	69.6	89.8	79.7	87.3	100.0	21.8	60.9	78.0
8	CXR Fact Encoder (T+C)	0.942	46.0	72.2	59.1	41.1	95.9	68.5	75.4	97.9	12.7	55.3	62.6
9	CXR Fact Encoder (T+R)	0.651	82.5	54.1	68.3	70.6	91.8	81.2	86.0	99.3	17.3	58.3	78.5
10	CXR Fact Encoder (T+SD)	0.629	71.0	48.4	59.7	78.4	79.6	79.0	81.9	99.3	13.2	56.2	70.3
11	CXR Fact Encoder (T+EC)	0.362	95.1	89.5	92.3	98.0	81.6	89.8	94.9	97.2	75.5	86.3	92.9
12	CXR Fact Encoder (T+EC+NLI)	0.313	96.0	89.9	93.0	99.0	89.8	94.4	94.9	98.6	71.4	85.0	94.4
13	CXR Fact Encoder (T+C+EC+NLI+ER)	0.491	95.6	83.0	89.3	96.1	81.6	88.9	93.1	100.0	61.4	80.7	94.9
14	CXR Fact Encoder (T+C+EC+NLI+SD)	0.512	95.8	81.8	88.8	95.1	83.7	89.4	93.5	100.0	50.9	75.5	94.9

Table 3: RadNLI test set accuracy results. Results for CXR-BERT, IFCC, PTUnifier and DoT5 are from the original papers.

ID	Text Model	Test Accuracy
1	CXR-BERT (Boecking et al., 2022)	65.2
2	IFCC (Miura et al., 2021)	77.8
3	PTUnifier (Chen et al., 2023a)	80.0
4	DoT5 (Liu et al., 2023a)	82.1
4	CXR Fact Encoder (T+EC+NLI)	71.3
5	CXR Fact Encoder (T+C+EC+NLI+ER)	75.6
6	CXR Fact Encoder (T+C+EC+NLI+SD)	78.1
7	CXR Fact Encoder (NLI fine-tuning)	79.8
8	GPT-4 + prompt engineering (see Fig. 21)	82.3

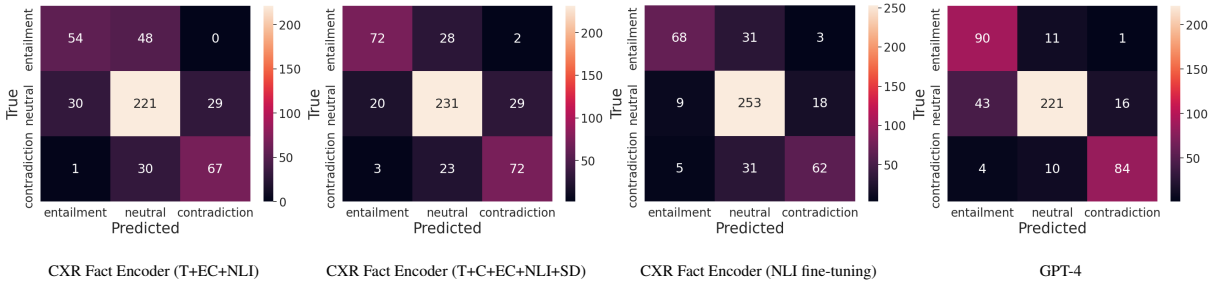


Figure 4: RadNLI test set confusion matrices

Table 4: Template-based report generation metrics on MIMIC-CXR test set for different label extraction methods. Notation: f denotes number of fact clusters, a denotes number of anatomical location clusters, and n denotes the maximum number of labels (only the n most frequent labels are kept). For *CXR Fact Encoder Score* with use CXR Fact Encoder (T+C+EC+NLI+SD) with a threshold of 0.7.

ID	Label Extraction Method	CXR Fact Encoder Score				RadGraph			CheXpert		CheXbert		BERTScore			BLEU	CIDEr-D	ROUGE-L	METEOR	
		F1	P	R	Sim	F1	P	R	Acc	F1	Acc	F1	F1	P	R					
1	CheXpert labeler (Irvin et al., 2019)	0.451	0.671	0.371	0.661	0.066	0.121	0.159	0.106	0.999	0.990	0.970	0.854	0.849	0.860	0.838	0.056	0.023	0.123	0.179
2	CheXbert (Smit et al., 2020)	0.454	0.677	0.371	0.664	0.067	0.122	0.161	0.107	0.974	0.921	0.992	0.907	0.849	0.860	0.838	0.056	0.023	0.123	0.179
3	Chest ImaGenome (Wu et al., 2021)	0.506	0.470	0.603	0.687	0.051	0.095	0.065	0.220	0.869	0.693	0.874	0.751	0.811	0.801	0.822	0.029	0.002	0.086	0.170
4	CXR Fact Encoder (f=200, a=50, n=1000)	0.831	0.840	0.826	0.833	0.140	0.241	0.287	0.214	0.867	0.633	0.863	0.671	0.865	0.878	0.853	0.088	0.033	0.189	0.240
5	CXR Fact Encoder (f=1000, a=300, n=10000)	0.932	0.939	0.928	0.897	0.186	0.307	0.342	0.287	0.885	0.686	0.909	0.747	0.875	0.888	0.863	0.116	0.070	0.223	0.290
6	CXR Fact Encoder (f=10000, a=300, n=50000)	0.974	0.983	0.966	0.943	0.268	0.414	0.444	0.398	0.937	0.826	0.944	0.844	0.890	0.901	0.880	0.164	0.138	0.289	0.364
7	CXR Fact Encoder (all facts)	0.982	0.993	0.974	0.986	0.644	0.776	0.799	0.768	0.986	0.964	0.979	0.946	0.927	0.939	0.916	0.366	0.555	0.523	0.630

CXR Fact Encoder (*all facts*) uses all the facts from ChatGPT/T5 without clustering, yielding the highest scores. This underscores the efficacy of the fact extraction process. Interestingly, *CXR Fact Encoder Score* suggests Chest ImaGenome surpasses CheXpert labeler and CheXbert in recall and F1 score but lags in precision. Yet, all three baseline methods are far from fully capturing the entire report information, a conclusion that is also supported by the RadGraph metrics, potentially due to their rigid annotation standards.

7 Conclusions

We have presented a novel two-stage system for the extraction and encoding of the factual information in radiology reports. The fact extraction stage can be effectively implemented by leveraging LLMs (ChatGPT and T5). For fact encoding, we have pre-

sented CXR Fact Encoder, a variant of CXR-BERT-specialized (Boecking et al., 2022) fine-tuned via multitask learning, with tasks like triplet ranking, quadruplet loss, natural language inference, sentence classification, sentence decoding and entity and relation extraction. In several of these tasks we leverage ChatGPT and T5 for added supervision, complementing expert-annotated datasets. The evaluations support the efficacy of our system. In addition, we release *CXR Fact Encoder Score*, a new radiology text generation evaluation metric that leverages the two stages of our system. We hope our work may inspire research towards better fact extraction and representation, improved LLM use, more advanced training protocols, and broader applications to downstream tasks such as image-based fact classification, fact visual grounding, VQA, report generation and summarization.

8 Limitations and Future Work

One significant limitation of our study is the absence of a thorough assessment by domain experts, such as radiologists, on both the prompts and the outputs generated by ChatGPT. While we diligently iterated the prompts and manually inspected the outputs on multiple examples, the ideal method would involve radiologists in the prompt engineering process, complemented by stringent evaluation protocols. This would ensure the most effective prompt strategies for the radiology field. Given this, we believe there’s untapped potential in utilizing LLMs more effectively for tasks like data augmentation, information extraction, and supervision generation. In this paper we’ve only scratched the surface of what is possible with these technologies.

Building on the earlier point, we see substantial potential for refining the triplet sampling heuristics outlined in Section 5.2. Involving radiologists in the heuristic design and validation of the generated triplets could be beneficial. Additionally, optimizing the use of LLMs with better prompts for triplet sampling and incorporating superior auxiliary embeddings could further enhance our approach.

Another significant limitation of our work is the omission of chest X-ray images paired with the reports. While our tests show advancements using just text, we recognize the critical value of visual data. Thus, we’re keen on exploring how CXR Fact Encoder can integrate image information within a multimodal framework. This could enhance tasks like image-driven report generation, VQA, and visual grounding of facts, to name a few. Exploring these avenues will be a primary focus in our subsequent research.

In this paper, our emphasis was on extracting facts from the *findings* and *impression* sections of a report. Yet, sections like *comparison*, *indication*, and *history* offer deeper insights and context about the patient. Expanding our fact extraction to encompass these sections and investigating how this broader patient information can be utilized to bolster downstream models’ performance is also an important avenue for future research.

Lastly, we acknowledge that our fact extraction algorithm faces a technical constraint: it extracts facts sentence-by-sentence, based on the `sent_tokenize` function from the NLTK library. This method could falter when a fact spans multiple sentences connected through co-reference. While such

occurrences are relatively uncommon in our observations, a deeper exploration of this linguistic aspect could guide the development of a more refined fact extraction mechanism that overcomes this challenge.

References

- Lisa C Adams, Daniel Truhn, Felix Busch, Avan Kader, Stefan M Niehues, Marcus R Makowski, and Keno K Bresslem. 2023. Leveraging gpt-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology*, 307(4):e230725.
- Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. ACL.
- Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision-language processing. In *European conference on computer vision*, pages 1–21. Springer.
- Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2019. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv:1901.07441*.
- Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. 2023a. Towards unifying medical vision-and-language pre-training via soft prompts. *arXiv preprint arXiv:2302.08958*.
- Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer.

764	Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2023b. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 469–484, Toronto, Canada. Association for Computational Linguistics.	822
765		823
766		
767		
768		
769		
770		
771		
772	Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards . In <i>Findings of the Association for Computational Linguistics: EMNLP 2022</i> , pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	
773		
774		
775		
776		
777		
778		
779		
780	Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. <i>Journal of the American Medical Informatics Association</i> , 23(2):304–310.	
781		
782		
783		
784		
785		
786		
787	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding . In <i>Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)</i> , pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.	
788		
789		
790		
791		
792		
793		
794		
795		
796	Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training . In <i>ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)</i> , volume 325 of <i>Frontiers in Artificial Intelligence and Applications</i> , pages 2006–2013. IOS Press.	
797		
798		
799		
800		
801		
802		
803		
804		
805		
806	Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing .	
807		
808		
809		
810		
811	Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, et al. 2023. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events . <i>arXiv preprint arXiv:2307.06439</i> .	
812		
813		
814		
815		
816		
817	Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison . In <i>Proceedings of the AAAI conference on artificial intelligence</i> , volume 33, pages 590–597.	824
818		825
819		826
820		827
821		828
		829
		830
		831
		832
		833
		834
		835
		836
		837
		838
		839
		840
		841
		842
		843
		844
		845
		846
		847
		848
		849
		850
		851
		852
		853
		854
		855
		856
		857
		858
		859
		860
		861
		862
		863
		864
		865
		866
		867
		868
		869
		870
		871
		872
		873
		874
		875
		876
		877

878	an iterative optimizing framework for radiology report summarization with chatgpt. <i>arXiv preprint arXiv:2304.08448</i> .	3982–3992, Hong Kong, China. Association for Computational Linguistics.	935 936
881	Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andia, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. <i>ACM Computing Surveys (CSUR)</i> , 54(10s):1–40.	Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain .	937 938 939
884	Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation . In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 5288–5304, Online. Association for Computational Linguistics.	Yucheng Shi, Hehuan Ma, Wenliang Zhong, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. 2023. Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. <i>arXiv preprint arXiv:2305.03513</i> .	940 941 942 943 944
887	Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. <i>IEEE Journal of Biomedical and Health Informatics</i> , 26(12):6070–6080.	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 . <i>CoRR</i> , abs/2004.09167.	945 946 947 948 949
895	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In <i>Proc of the 40th Annual Meeting of the ACL</i> , pages 311–318, Philadelphia, Pennsylvania, USA. ACL, ACL.	Tanveer Syeda-Mahmood, Ken CL Wong, Yaniv Gur, Joy T Wu, Ashutosh Jadhav, Satyananda Kashyap, Alexandros Karagyris, Anup Pillai, Arjun Sharma, Ali Bin Syed, et al. 2020. Chest x-ray report generation through fine-grained label learning. In <i>Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23</i> , pages 561–571. Springer.	950 951 952 953 954 955 956 957 958
901	Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.	Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 7433–7442.	959 960 961 962 963
906	Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. 2021. Clinically correct report generation from chest x-rays using templates. In <i>Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12</i> , pages 654–663. Springer.	Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In <i>Proc of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)</i> , pages 4566–4575.	964 965 966 967 968
910	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 <i>International conference on machine learning</i> , pages 8748–8763. PMLR.	David Wadden, Ulme Wennberg, Yi Luan, and Hananeh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.	969 970 971 972 973 974 975 976 977
917	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. <i>The Journal of Machine Learning Research</i> , 21(1):5485–5551.	Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. <i>Advances in Neural Information Processing Systems</i> , 35:33536–33549.	978 979 980 981 982
923	Nils Reimers and Iryna Gurevych. 2019. SentenceBERT: Sentence embeddings using Siamese BERT-networks . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages	Joy T Wu, Nkechinyere Agu, Ismini Lourentzou, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. 2021. Chest imagenome dataset for clinical reasoning . In <i>Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks</i> , volume 1. Curran.	983 984 985 986 987 988 989 990 991

992 Yue Yang, Artemis Panagopoulou, Shenghao Zhou,
 993 Daniel Jin, Chris Callison-Burch, and Mark Yatskar.
 994 2023. Language in a bottle: Language model guided
 995 concept bottlenecks for interpretable image classifica-
 996 tion. In *Proceedings of the IEEE/CVF Conference*
 997 *on Computer Vision and Pattern Recognition*, pages
 998 19187–19197.

999 Michihiro Yasunaga, Jure Leskovec, and Percy Liang.
 1000 2022. Linkbert: Pretraining language models with
 1001 document links. In *Association for Computational*
 1002 *Linguistics (ACL)*.

1003 Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan,
 1004 Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser
 1005 Ururahy Nunes Fonseca, Henrique Min Ho Lee,
 1006 Zahra Shakeri Hossein Abad, Andrew Y Ng, et al.
 1007 2022. Evaluating progress in automatic chest x-ray
 1008 radiology report generation. *medRxiv*, pages 2022–
 1009 08.

1010 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q.
 1011 Weinberger, and Yoav Artzi. 2020a. *Bertscore: Eval-*
 1012 *uating text generation with bert*. In *International*
 1013 *Conference on Learning Representations*.

1014 Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D.
 1015 Manning, and Curtis Langlotz. 2020b. *Optimizing*
 1016 *the factual correctness of a summary: A study of*
 1017 *summarizing radiology reports*. In *Proceedings of*
 1018 *the 58th Annual Meeting of the Association for Com-*
 1019 *putational Linguistics*, pages 5108–5120, Online. As-
 1020 sociation for Computational Linguistics.

1021 A Appendix

1022 A.1 CXR Fact Encoder’s Tasks Details

1023 CXR Fact Encoder is a fine-tuned version of
 1024 CXR-BERT-specialized, which can be downloaded
 1025 from [https://huggingface.co/microsoft/](https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-specialized)
 1026 [BiomedVLP-CXR-BERT-specialized](https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-specialized). One of the
 1027 tasks we explore for model fine-tuning is sentence
 1028 ranking via triplet loss. Figure 5 illustrates this
 1029 task. Concretely, we forward 3 sentences (anchor,
 1030 positive, negative) through CXR-BERT-specialized
 1031 with weight sharing, obtaining three vectors a ,
 1032 b , and c each of dimension 128, and compute
 1033 $\Delta\text{sim}(a, p, n) = a \cdot p - a \cdot n$. This serves as the
 1034 input logit for a binary cross-entropy loss.

1035 A second group of tasks are classification tasks
 1036 (Figure 6). These include category (5 classes:
 1037 *anatomical finding, disease, technical assessment,*
 1038 *tubes and lines and device*), health status (4 classes:
 1039 i.e., *normal, abnormal, ambiguous, unknown*),
 1040 comparison status (15 classes, see Figure 15),
 1041 Chest ImaGenome observations (74 classes, see
 1042 Figure 16) and anatomical locations (38 classes,
 1043 see Figure 17). Category, Health Status and Com-
 1044 parison Status are single-label multi-class classi-

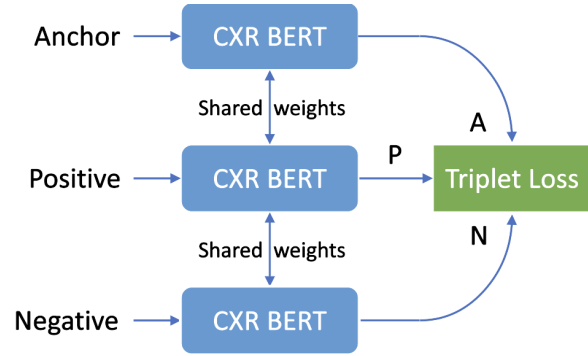


Figure 5: Triplet loss (T)

1045 fication tasks, whereas Chest ImaGenome obser-
 1046 vations and anatomical locations are multi-label
 1047 multi-class classification tasks. Implementing these
 1048 tasks require attaching fully connected heads on
 1049 top of CXR-BERT-specialized’s built-in projection
 1050 layer in order to perform the classification.

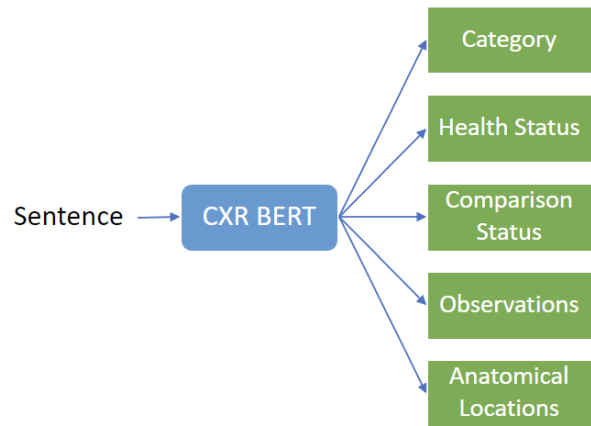


Figure 6: Sentence classification (C)

1051 Another task is sentence decoding (Figure 7).
 1052 We attach a lightweight, shallow Transformer De-
 1053 coder to CXR-BERT-specialized’s projection layer
 1054 in order to generate back the original sentence. This
 1055 can be viewed a sort of text autoencoder, forcing
 1056 the projection layer to capture as much information
 1057 as possible of the input sentence to facilitate the
 1058 reconstruction of the sentence by the Transformer
 1059 Decoder. We use a Transformer Decoder with em-
 1060 bedding, hidden and feedforward dimension 256,
 1061 only one self-attention head and only one layer.



Figure 7: Sentence decoding (SD)

The next task is what we refer to as entailment/-

contradiction quadruplet loss (Figure 8). The goal of this task is to promote a generalized separation of entailment and contradiction sentence pairs in the latent space, by randomly sampling entailment and contradiction pairs and requiring that the entailment pair have greater similarity than the contradiction pair. This loss was crucial to achieve state-of-the-art results in Table 2.

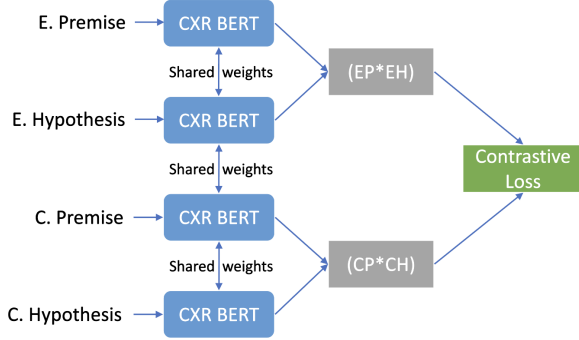


Figure 8: Entailment/contradiction quadruplet loss (EC)

For NLI, we adopt an approach similar to that of SBERT (Reimers and Gurevych, 2019), by concatenating the embeddings of the premise, hypothesis and their element-wise multiplication, followed by a softmax layer for NLI classification (see Figure 9).

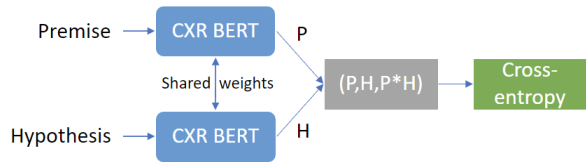


Figure 9: Natural language inference (NLI)

Lastly, for entity and relation extraction we augment CXR-BERT-specialized with the layers proposed by SpERT (Eberts and Ulges, 2020). This adaptation was relatively straightforward, since the authors of SpERT released an implementation (<https://github.com/lavis-nlp/spert/>) that is compatible with Huggingface models like CXR-BERT-specialized.

A.2 Label Extraction Details

In order to extract labels with our two-stage system, we set as a goal to select representative facts that would be assigned as labels to a given report. For that, we run K-Medoids clustering⁵ over

⁵https://scikit-learn-extra.readthedocs.io/en/stable/generated/sklearn_extra.cluster.KMedoids.html

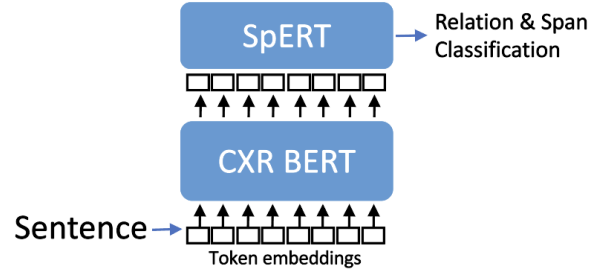


Figure 10: Entity and relation extraction (ER) with SpERT

all fact sentence embeddings (2,212,958 counting paraphrases) with F cluster centers, and K-Medoids clustering for all anatomical location sentences (296,434 counting paraphrases) with A cluster centers. Then, for each fact f extracted from a report, we assign to it the closest fact cluster center and the closest anatomical location cluster center (if the fact has an anatomical location). This produces labels of the form $(fact_cluster_id, anatomy_cluster_id)$ or just $fact_cluster_id$. Then we count the frequency of these labels and keep the N most frequent. For $fact_cluster_id$ labels, we simply choose the fact that K-Medoids clustering determined as the cluster center. For $(fact_cluster_id, anatomy_cluster_id)$ labels, we go over all the facts producing the same pair and choose the fact that minimizes the sum of the inverse of the frequency of each word as a way of estimating the rareness of a sentence (i.e., we pick the least "rare" fact). Please refer to Table 5 to see examples of template-based reports built in this way, along with examples for CheXpert labeler, CheXbert and Chest ImaGenome.

A.3 Implementation Details

All of our experiments are implemented using Python 3.10.10 with PyTorch version 1.13.1+cu117 (Paszke et al., 2017). All experiments are conducted on a computing node equipped with a 20-core Intel(R) Core(TM) i9-9900X CPU @ 3.50GHz, two NVIDIA GPUs - one GeForce RTX 2080 Ti with 11GB memory and one GeForce RTX 3090 with 24GB memory. The system is complemented by 125GB of RAM.

We implement multitask learning for CXR Fact Encoder through gradient accumulation. This is achieved by multiple model forwards, each fed by interleaved dataloaders for different tasks. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a cyclic exponential learning rate

Table 5: Examples of template-based generated reports for different label extraction algorithms.

Ground-truth Report	CXR Fact Encoder (all facts)	CXR Fact Encoder (f=10000, a=300, n=50000)	CXR Fact Encoder (f=1000, a=300, n=10000)
<p>New PICC line on the right is projecting with its tip somewhere in the mediastinum. Appears to cross the midline, there is concern for potential arterial location. The initial line concerns were communicated over the telephone at the time of the wet read. Repeat PA and lateral radiograph, taken approximately an hour after the radiograph demonstrated the PICC line in the mid SVC. Potential small right pleural effusion. Stable moderate cardiomegaly.</p>	<p>new PICC line on the right. tip of the PICC line in the mediastinum. potential arterial location crossing the midline. PICC line in the mid SVC. potential small right pleural effusion. stable moderate cardiomegaly</p> <p>CXR Fact Encoder Score (sim): 1.000 RadGraph F1: 0.796 CheXpert Acc: 1.0 CheXbert Acc: 1.0</p>	<p>new right PICC. tip of the PICC line in the mediastinum. catheter crosses midline. PICC in mid SVC. likely right effusion. mild to moderate cardiomegaly unchanged</p> <p>CXR Fact Encoder Score (sim): 0.954 RadGraph F1: 0.539 CheXpert Acc: 1.0 CheXbert Acc: 1.0</p>	<p>new right PICC. tip of the PICC line in the mediastinum. projecting over the midline. PICC in mid SVC. small right effusion. unchanged evidence of cardiomegaly</p> <p>CXR Fact Encoder Score (sim): 0.934 RadGraph F1: 0.455 CheXpert Acc: 1.0 CheXbert Acc: 1.0</p>
CXR Fact Encoder (f=200, a=50, n=1000)	CheXbert	CheXpert labeler	Chest ImaGenome
<p>right pleural tube. The right PICC line terminates in the middle of the SVC. femoral catheter. tip of the mid SVC. right effusion. stable cardiomegaly is unchanged</p> <p>CXR Fact Encoder Score (sim): 0.794 RadGraph F1: 0.460 CheXpert Acc: 0.929 CheXbert Acc: 0.929</p>	<p>the heart is enlarged. the cardiomeastinal silhouette is enlarged. no focal consolidation. the lungs are free of focal airspace disease. no atelectasis. a device is seen. pleural effusion is seen. no fibrosis. no pneumonia. no pneumothorax is seen. no pulmonary edema. no pulmonary nodules or mass lesions identified. no fracture is seen</p> <p>CXR Fact Encoder Score (sim): 0.517 RadGraph F1: 0.021 CheXpert Acc: 1.0 CheXbert Acc: 1.0</p>	<p>the heart is enlarged. the cardiomeastinal silhouette is enlarged. no focal consolidation. the lungs are free of focal airspace disease. no atelectasis. a device is seen. pleural effusion is seen. no fibrosis. no pneumonia. no pneumothorax is seen. no pulmonary edema. no pulmonary nodules or mass lesions identified. no fracture is seen</p> <p>CXR Fact Encoder Score (sim): 0.517 RadGraph F1: 0.021 CheXpert Acc: 1.0 CheXbert Acc: 1.0</p>	<p>enlarged cardiac silhouette in cardiac silhouette. abnormal cardiac silhouette. picc in left shoulder. picc in mediastinum. lung opacity in right costophrenic angle. pleural effusion in right costophrenic angle. abnormal right costophrenic angle. lung opacity in right lung. pleural effusion in right lung. abnormal right lung. picc in right shoulder. picc in svc. enlarged cardiac silhouette. lung opacity. pleural effusion. picc</p> <p>CXR Fact Encoder Score (sim): 0.647 RadGraph F1: 0.103 CheXpert Acc: 0.929 CheXbert Acc: 0.929</p>

1129 that varies from $8e-5$ to $1e-6$ over 8 epochs. Here,
1130 an epoch consists of roughly 800 batches. Typi-
1131 cally, our experiments run for 10-12 hours, after
1132 which we observe no significant gains in validation
1133 metrics.

1134 **A.4 ChatGPT prompts**

Playground

SYSTEM

Relevant facts:

1. observations of abnormalities
2. observations of diseases
3. observations of strange visual patterns
4. observations of devices
5. observations of foreign bodies
6. observations of specific anatomical regions that look normal or healthy
7. absences of abnormalities (usually expressed with a negation)
8. comparisons with respect to a previous study (something changed or remained the same)

Task:

Given a sentence taken from a chest x-ray report, generate a JSON list of relevant facts. Each fact should be about one observation. If a sentence mentions multiple observations, each observation should be extracted as a separate fact. Each fact should include the anatomical location where it was observed. If multiple facts occur in the same location, repeat the location in each fact.

If no relevant facts are mentioned, return [] (an empty array).

Examples:

Opacity and density in the right lobe

```
[  
  "opacity in the right lobe",  
  "density in the right lobe"  
]
```

Lungs are well inflated without evidence of focal airspace consolidation to suggest pneumonia.

```
[  
  "well inflated lungs",  
  "lungs without evidence of focal airspace consolidation",  
  "lungs without evidence of pneumonia"  
]
```

SYSTEM

```
consolidation",  
"lungs without evidence of pneumonia"  
]  
  
Taken together, compared with less than 1 hr earlier, the findings are suggestive of worsening of CHF, with new or significantly increased left greater right pleural effusions and underlying bibasilar collapse and/or consolidation, particularly on the left.  
[  
  "worsening of CHF",  
  "new or significantly increased left pleural effusions",  
  "new or significantly increased right pleural effusions",  
  "underlying bibasilar collapse on the left",  
  "underlying consolidation on the left",  
]  
  
No acute cardiopulmonary abnormality  
[  
  "no acute cardiopulmonary abnormality"  
]
```

Your presets

USER There is no pneumothorax, focal consolidation, or pleural effusion

ASSISTANT

```
[  
  "no pneumothorax",  
  "no focal consolidation",  
  "no pleural effusion"  
]
```

Mode:

Model:

Temperature: 0

Maximum length: 512

Stop sequences: Enter sequence and press Tab

Top P: 1

Frequency penalty: 0

Presence penalty: 0

API and Playground requests will not be used to train our models. [Learn more](#)

Figure 11: Screenshot of OpenAI's playground web interface showing the prompt used to extract facts from chest X-ray report sentences. The screenshot was manipulated in order to display the full instructions given to ChatGPT.

SYSTEM

Given a fact extracted from a Chest X-ray report, output a JSON array of strings. Each string must be a paraphrased fact, with similar semantics, expressing the same finding/diagnosis. These paraphrases will be used to train a fact embedding with triplet loss.

Generate multiple examples (at least 14), covering a wide variety and diversity of radiological terms, synonyms, abbreviations and expressions that radiologists commonly use to communicate the same idea.

Output format:

```
[
  "paraphrase 1",
  "paraphrase 2",
  ...
  "paraphrase 14"
]
```

USER osteoporosis

ASSISTANT [

- "Bone thinning is observed",
- "Signs of bone loss are present",
- "Indications of decreased bone density",
- "Evidence of porous bones",
- "Presence of brittle bones",
- "Manifestations of bone weakening",
- "Indications of bone fragility",
- "Signs of demineralized bones",
- "Presence of low bone mass",
- "Manifestations of bone softening",
- "Evidence of bone rarefaction",
- "Indications of skeletal fragility",
- "Signs of reduced bone strength",
- "Presence of bone tissue loss"

]

➤ Add message

Submit ↻

Mode
Chat ▾

Model
gpt-4-0613 ▾

Temperature 0

Maximum length 512

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

🔒 API and Playground requests will not be used to train our models. [Learn more](#)

[Give us feedback](#)

Figure 12: Fact to paraphrases

SYSTEM

Given a sentence referring to an anatomical location in the context of a chest X-ray, output a JSON array of strings paraphrasing it, covering a wide diverse range of terminology, synonyms and abbreviations that radiologists commonly use to express the same idea in a chest X-ray report.

Examples:

```
gastroesophageal junction
[
"cardia",
"cardia of the stomach",
"cardiac orifice",
"esophagogastric junction",
"GE junction",
"esophagogastric junction",
"esophagogastric junctional region",
"gastroesophageal interface",
"gastroesophageal transition zone",
"lower esophageal junction",
"esophagogastric transition",
"stomach-esophagus interface",
"junction between stomach and
```

SYSTEM

```
"junction between stomach and
esophagus",
"cardioesophageal junction",
"junction of the esophagus and stomach",
"transition between the esophagus and stomach",
"point where the esophagus meets the stomach",
"Z-line",
"distal end of the esophagus",
"proximal start of the gastric cardia"
]

costodiaphragmatic recess
[
"costophrenic angle",
"costophrenic recess",
"costophrenic sulcus",
"costophrenic notch",
"costophrenic sinus",
"phrenicocostal sinus",
"diaphragm-chest wall interface",
"junction of diaphragm and chest wall",
"area where the diaphragm meets the ribs",
"junction between the diaphragm and ribs"
```

SYSTEM

```
"costal margin of the diaphragm",
"angle formed by diaphragm and rib cage"
]

R>L
[
"right greater than left"
]

base
[
"lung base",
"base of the lung",
"lower lung area",
"basal lung region",
"basal portion of the lung",
"lung's basal part",
"bottom part of the lung",
"lower region of the lung",
"pulmonary base",
"base of the pulmonary structure",
"lower zone of the lung",
"lower pulmonary region",
"lower segment of the lung"
]
```

USER cardiomediastinal silhouette

ASSISTANT [

```
"heart and mediastinal contour",
"cardiac and mediastinal silhouette",
"outline of the heart and mediastinum",
"cardiac and mediastinal shadow",
"cardiopulmonary silhouette",
"heart and mediastinal structure",
"cardiac and mediastinal shape",
"cardiac and mediastinal border",
"cardiac and mediastinal profile",
"cardiac and mediastinal image",
"cardiac and mediastinal area",
"cardiac and mediastinal configuration",
"cardiac and mediastinal pattern",
"cardiac and mediastinal appearance",
"cardiac and mediastinal contouring",
"cardiac and mediastinal form",
"cardiac and mediastinal outline",
"cardiac and mediastinal figure",
"cardiac and mediastinal boundary"
]
```

[Submit](#) ↻

Mode

Model

Temperature 0

Maximum length 512

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

API and Playground requests will not be used to train our models. [Learn more](#)

Figure 13: Anatomy location to paraphrases

SYSTEM

Given a medical fact, output a JSON object with 7 fields:

- "anatomical location"
- "detailed observation"
- "short observation"
- "category"
- "health status"
- "prev_study_comparison?"
- "comparison status"

Definitions:

- "anatomical location" means the anatomical location of the observation. If not given, it should be "".
- "detailed observation" means what was observed, with all the details provided (but excluding location, unless the location itself is the main observation).
- "short observation" should be a brief summary of the previous field, without extra details, but logically consistent.
- "category" can be "anatomical finding", "disease", "technical assessment", "tubes and lines" or "device".

SYSTEM

5. "health status" can be "normal", "abnormal", "ambiguous" or "unknown".
 "normal" means healthy.
 "abnormal" means unhealthy.
 "ambiguous" means it is not clear whether it is healthy or unhealthy, leaving room for interpretation.
 "unknown" means the observation provides no information about abnormality/normality.

6. "prev_study_comparison?" is a boolean field. Can be either "yes" or "no". If the fact is comparing the current study with a previous study (something improved, worsened, changed or remained the same), return "yes". Otherwise, return "no".

7. "comparison status" can be "resolved", "new", "improved", "worsened", "larger", "smaller", "displaced", or "". If no comparison with a previous study is made, the default is "" (empty).

Examples:

```
small to moderate size left pleural effusion
{
  "anatomical location": "left",
  "detailed observation": "small to moderate size pleural effusion",
  "short observation": "pleural effusion",
```

SYSTEM

```
"category": "anatomical finding",
"health status": "abnormal",
"prev_study_comparison?": "no",
"comparison status": ""
}

new finding of mass in the abdomen
{
  "anatomical location": "abdomen",
  "detailed observation": "mass",
  "short observation": "mass",
  "category": "anatomical finding",
  "health status": "abnormal",
  "prev_study_comparison?": "yes",
  "comparison status": "new"
}

clear lung fields
{
  "anatomical location": "lung fields",
  "detailed observation": "clear lung fields",
  "short observation": "clear lung fields",
  "category": "anatomical finding",
  "health status": "normal",
  "prev_study_comparison?": "no",
  "comparison status": ""
}

no ill-definition of right hemidiaphragm
```

USER bilateral effusions equivocally getting worse

ASSISTANT {
 "anatomical location": "bilateral",
 "detailed observation": "effusions equivocally getting worse",
 "short observation": "worsening effusions",
 "category": "anatomical finding",
 "health status": "abnormal",
 "prev_study_comparison?": "yes",
 "comparison status": "worsened"
 }

Mode
Chat

Model
gpt-4-0613

Temperature 0

Maximum length 512

Stop sequences
Enter sequence and press Tab

Figure 14: Fact to metadata

SYSTEM

Given a statement extracted from a chest X-ray report, output a comparison category. The category must be one item of the following list:

0. no comparison
1. new finding
2. resolved
3. improved
4. worsened
5. progressed
6. reappeared
7. larger
8. smaller
9. increase
10. decrease
11. position changed
12. stable/unchanged
13. unclear comparison
14. other

Examples:

|
right pleural effusion similar in size
12. stable/unchanged

pleural density on the left is somewhat increased
9. increase

new bibasal consolidations on the left

SYSTEM

6. reappeared
7. larger
8. smaller
9. increase
10. decrease
11. position changed
12. stable/unchanged
13. unclear comparison
14. other

Examples:

|
right pleural effusion similar in size
12. stable/unchanged

pleural density on the left is somewhat increased
9. increase

new bibasal consolidations on the left
1. new finding

small foci of opacity in the left mid lung
0. no comparison

no ill-definition of right hemidiaphragm
0. no comparison

heart within upper limits of normal in size
0. no comparison

USER bilateral effusions equivocally getting worse

ASSISTANT 4. worsened

⊕ Add message

Mode

Chat

Model

gpt-4-0613

Temperature 0

Figure 15: Fact to comparison status

SYSTEM

Instructions:

Apply an annotation standard of observations to raw phrases extracted from chest X-ray reports. The standard considers the following observations:

- airspace opacity
- atelectasis
- bone lesion
- bronchiectasis
- calcified nodule
- clavicle fracture
- consolidation
- costophrenic angle blunting
- cyst/bullae
- diaphragmatic eventration (benign)
- elevated hemidiaphragm
- enlarged cardiac silhouette
- enlarged hilum
- hernia
- hydropneumothorax
- hyperaeration
- increased reticular markings/ijq pattern
- infiltration
- linear/patchy atelectasis
- lobar/segmental collapse
- lung lesion
- lung opacity
- mass/nodule (not otherwise specified)

SYSTEM

- mediastinal displacement
- mediastinal widening
- multiple masses/nodules
- pleural effusion
- pleural/parenchymal scarring
- pneumomediastinum
- pneumothorax
- pulmonary edema/hazy opacity
- rib fracture
- scoliosis
- shoulder osteoarthritis
- spinal degenerative changes
- spinal fracture
- sub-diaphragmatic air
- subcutaneous air
- superior mediastinal mass/enlargement
- tortuous aorta
- vascular calcification
- vascular congestion
- vascular redistribution
- aortic graft/repair
- cabg grafts
- cardiac pacer and wires
- prosthetic valve
- alveolar hemorrhage
- aspiration
- coppd/emphysema
- fluid overload/heart failure
- goiter

SYSTEM

- granulomatous disease
- interstitial lung disease
- lung cancer
- pericardial effusion
- pneumonia
- artifact
- breast/nipple shadows
- low lung volumes
- rotated
- skin fold
- chest port
- chest tube
- endotracheal tube
- enteric tube
- ij line
- intra-aortic balloon pump
- mediastinal drain
- piccc
- pigtail catheter
- subclavian line
- swan-ganz catheter
- tracheostomy tube

Output format:

JSON array of strings

Rules:

Only output the observations from the standard that best match or are supported by the phrase. If none match the phrase, output an empty array ([]).

Example:

Increased bibasilar opacities likely representing atelectasis, but cannot exclude aspiration or pneumonia in the correct clinical setting

```
[
  "lung opacity",
  "atelectasis",
  "aspiration",
  "pneumonia"
]
```

SYSTEM

- intra-aortic balloon pump
- mediastinal drain
- piccc
- pigtail catheter
- subclavian line
- swan-ganz catheter
- tracheostomy tube

Output format:

JSON array of strings

Rules:

Only output the observations from the standard that best match or are supported by the phrase. If none match the phrase, output an empty array ([]).

Example:

Increased bibasilar opacities likely representing atelectasis, but cannot exclude aspiration or pneumonia in the correct clinical setting

```
[
  "lung opacity",
  "atelectasis",
  "aspiration",
  "pneumonia"
]
```

USER dual chamber pacemaker

ASSISTANT [

"cardiac pacer and wires"

]

⊕ Add message

Mode

Chat

Model

gpt-4-0613

Temperature 0

Maximum length 512

Figure 16: Fact to observations

SYSTEM

Instructions:

Apply an annotation standard of anatomical locations to raw phrases extracted from chest X-ray reports. The standard considers the following anatomical locations:

- abdomen
- aortic arch
- cardiac silhouette
- carina
- cavoatrial junction
- left apical zone
- left arm
- left breast
- left chest wall
- left clavicle
- left costophrenic angle
- left hemidiaphragm
- left hilar structures
- left lower lung zone
- left lung
- left mid lung zone
- left shoulder
- left upper lung zone
- mediastinum
- neck
- right apical zone

SYSTEM

- right arm
- right atrium
- right breast
- right chest wall
- right clavicle
- right costophrenic angle
- right hemidiaphragm
- right hilar structures
- right lower lung zone
- right lung
- right mid lung zone
- right shoulder
- right upper lung zone
- spine
- svc
- trachea
- upper mediastinum

Output format:

JSON array of strings

Rules:

Only output the anatomical locations from the standard that best match the given phrase. If none match the phrase, output an empty array ([]).

SYSTEM

Only output the anatomical locations from the standard that best match the given phrase. If none match the phrase, output an empty array ([]).

Examples:

cardiomediastinal silhouette

```
[
  "cardiac silhouette",
  "mediastinum"
]
```

A tracheostomy tube is in place

```
[
  "neck",
  "trachea"
]
```

bilateral pulmonary nodules

```
[
  "left lung",
  "right lung"
]
```

left IJV

```
[
  "neck"
]
```

USER dual chamber pacemaker

ASSISTANT [

"right atrium",

"right ventricle"

]

⊕ Add message

Mode

Chat

Model

gpt-4-0613

Temperature 0

Maximum length 512

Figure 17: Fact to anatomical locations

SYSTEM

Given an anchor, output a list of positives and a list of negatives.

The anchor will come from a chest X-ray report. The positives must be 6 paraphrases of the anchor, expressing the same idea with different terms and synonyms (cover a wide range of medical terminology).

The negatives must be 10 hard negatives, that is, similar to the anchor but semantically different, for example, by expressing a different diagnosis (if applicable), by referring to a different anatomical location (if applicable), etc.

Output format: a JSON object as follows

```
{
  "positives": [...],
  "negatives": [...]}

```

USER removal of endotracheal tube

ASSISTANT {

```
"positives": [
  "Endotracheal tube has been removed",
  "Extraction of the endotracheal tube",
  "Endotracheal tube has been taken out",
  "Endotracheal tube has been extracted",
  "Endotracheal tube removal has been performed",
  "Discontinuation of the endotracheal tube"
],
"negatives": [
  "Insertion of endotracheal tube",
  "Removal of nasogastric tube",
  "Removal of chest tube",
  "Endotracheal tube is in place",
  "Endotracheal tube repositioned",
  "Removal of central venous catheter",
  "Endotracheal tube is unchanged",
  "Removal of tracheostomy tube",
  "Endotracheal tube is malpositioned",
  "Removal of urinary catheter"
]
}
```

Mode
Chat

Model
gpt-4-0613

Temperature 0

Maximum length 512

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

API and Playground requests will not be used to train our models. [Learn more](#)

Figure 18: Hard triplets generated by ChatGPT

SYSTEM

Given a premise from a Chest X-ray report, output an entailment, a contradiction and a neutral sentence.

The entailment must strictly follow logically from the premise. The neutral can optionally be about a different chest X-ray observation. The output format must be a JSON as follows:

```
{
  "e": ...,
  "c": ...,
  "n": ...}

```

USER removal of endotracheal tube

ASSISTANT {

```
"e": "The patient no longer has an endotracheal tube.",
"c": "The endotracheal tube is still in place.",
"n": "The patient's lungs appear clear."
}
```

⊕ Add message

Mode
Chat

Model
gpt-4-0613

Temperature 0

Maximum length 512

Stop sequences
Enter sequence and press Tab

Figure 19: NLI Prompt 1

SYSTEM

You will receive a NLI example from a chest X-ray dataset with premise, hypothesis and label ("entailment", "contradiction" or "neutral"). Generate 10 new NLI examples with the same label and following a similar writing style as the example provided. They have to be about chest X-ray reports. Output the examples as a JSON array of objects. If the premise or the hypothesis (or both) are complex (i.e. they say several things), create challenging examples following a similar structure as the provided example that require a good understanding of logic in order to deduce the label.

Avoid verbosity and words such as "patient" or "X-ray". State observations succinctly.

USER {"P": "No pleural effusions or pneumothorax.", "H": "Previously seen pneumothorax is no longer visualized.", "L": "entailment"}

ASSISTANT [

```
{ "P": "No evidence of pulmonary edema or pneumonia.", "H": "The previously noted pneumonia is not present.", "L": "entailment" },
{ "P": "No signs of lung nodules or masses.", "H": "The lung mass seen before is not visible now.", "L": "entailment" },
{ "P": "No visible signs of cardiomegaly or mediastinal widening.", "H": "The previously observed cardiomegaly is not seen.", "L": "entailment" },
{ "P": "No signs of pulmonary embolism or pleural effusion.", "H": "The earlier detected pleural effusion is not visible.", "L": "entailment" },
{ "P": "No evidence of pneumothorax or lung collapse.", "H": "The previously seen lung collapse is not present.", "L": "entailment" },
{ "P": "No signs of rib fractures or lung contusions.", "H": "The rib fracture noted before is not visible now.", "L": "entailment" },
{ "P": "No visible signs of bronchiectasis or lung fibrosis.", "H": "The previously observed bronchiectasis is not seen.", "L": "entailment" },
{ "P": "No signs of lung cancer or tuberculosis.", "H": "The earlier detected tuberculosis is not visible.", "L": "entailment" },
{ "P": "No evidence of aortic dissection or aneurysm.", "H": "The previously seen aortic aneurysm is not present.", "L": "entailment" },
{ "P": "No signs of pericardial effusion or heart failure.", "H": "The previously observed pericardial effusion is not visible.", "L": "entailment" }
]
```

Mode
Chat

Model
gpt-4-0613

Temperature 0

Maximum length 512

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

API and Playground requests will not be used to train our models. [Learn more](#)

Figure 20: NLI Prompt 2

SYSTEM

Context: natural language inference.

Given a premise and a hypothesis, output "entailment", "contradiction", or "neutral".

Use "entailment" when the facts stated by the premise necessarily entail the truth of the hypothesis.

Use "contradiction" when premise and hypothesis are mutually exclusive/contradictory (both cannot be true at the same time).

Use "neutral", if there is no contradiction (premise and hypothesis are compatible), but the premise does not entail the hypothesis (it's possible for the premise to be true and the hypothesis still be false). In other words, use "neutral" when neither "entailment" nor "contradiction" adequately fit.

USER Premise: Lungs are well-expanded and clear without focal consolidation concerning for pneumonia. | Hypothesis: Lungs are hyperinflated but clear.

ASSISTANT Neutral

USER Enter a user message here.

⊕ Add message

Mode
Chat

Model
gpt-4-0613

Temperature 0

Maximum length 10

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

Figure 21: NLI Prompt 3

SYSTEM

Given a premise from a Chest X-ray report, output 5 statements that explicitly logically contradict the premise, that is, the premise and any generated statement cannot both be true at the same time. Include at least one sentence that is very similar to the premise but with a very slight difference, thus creating a logical contradiction.

For example:

- "There is mid thoracic dextroscoliosis" vs. "There is no mid thoracic dextroscoliosis"
- "Left basal consolidation has slightly improved" vs. "Left basal consolidation has slightly increased"

Output format: a JSON array of strings.

USER Lungs are well-expanded and clear without focal consolidation concerning for pneumonia.

ASSISTANT ["Lungs are not well-expanded and clear without focal consolidation concerning for pneumonia.", "Lungs are well-expanded but show focal consolidation concerning for pneumonia.", "Lungs are collapsed and clear without focal consolidation concerning for pneumonia.", "Lungs are well-expanded and clear with focal consolidation not concerning for pneumonia.", "Lungs are well-expanded and clear, but there is focal consolidation concerning for pneumonia."]

⊕ Add message

Mode
Chat

Model
gpt-4-0613

Temperature 0

Maximum length 500

Stop sequences
Enter sequence and press Tab

Top P 1

Figure 22: NLI Prompt 4