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

Comparison: Chest radiographs XXXX.
Indication: XXXX-year-old male, chest pain.
Findings: The cardiomediastinal 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)

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 example 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 API[1]. 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 library [2], 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 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 distil 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/

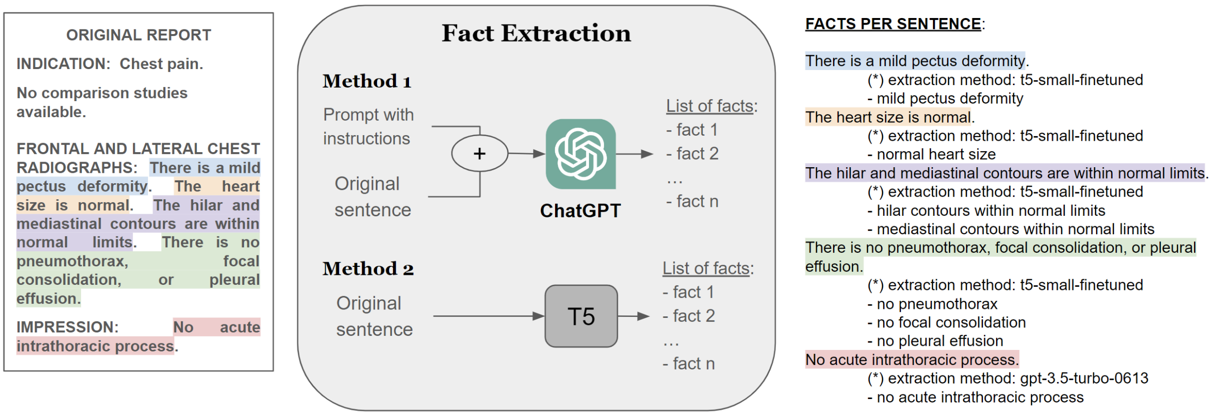


Figure 2: Fact extraction procedure for radiology reports.

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 hub[3]. 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 separation 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.

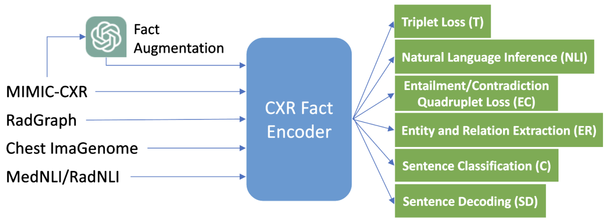


Figure 3: CXR Fact Encoder model.

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

[3] https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-specialized

frontal view images in MIMIC-CXR, Chest ImaGenome gives a scene graph connecting report observations to image anatomical locations via bounding boxes. This is achieved through a combined rule-based NLP and atlas-based bounding box detection process, backed by a CXR lexicon and ontology 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, facilitating the creation of a binary multi-label classification task for a text encoder to predict sentence-based observations and locations. 2. Chest ImaGenome introduces a radiologist-informed annotation 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 annotations 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 reports, each annotated by two board-certified radiologists, and 500 CheXpert reports annotated by DyGIE++. RadGraph is incorporated into our experiments due to its rich domain-specific annotations.

As we conduct experiments on NLI, we also leverage the datasets MedNLI (Romanov and Shivade, 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 annotations 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, encompassing fields like "anatomical location", "detailed observation", "short observation", "category", "health status" and "comparison status". The respective 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 ImaGenome dataset's annotation standards, as demonstrated in Figures 16 and 17. Notably, when adding metadata and Chest ImaGenome labels to facts, we adopt the approach detailed in Section 3: we selectively use ChatGPT for a subset and then train T5 for the remaining annotations. This method lets us expand Chest ImaGenome 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 sentences 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 computed as $\Delta\text{sim}(a, p, n) = \text{sim}(a, p) - \text{sim}(a, n)$ from their embeddings' dot product. By minimizing the binary cross-entropy loss, the encoder ensures closely related sentences are nearer and unrelated 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, lev(x, y) for the levenshtein string distance between them, and levsim(x, y) = 1 - lev(x, y) / max(len(x), len(y)). c(x) indicates 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 generated by ChatGPT and n is any other sentence (unless cos(E(a), E(p)) $<$ cos(E(a), E(n)) and lev(a, p) $>$ lev(a, n)).

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

Rule 3: Short observation, detailed observation and the original fact (and their paraphrases) 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 lev(a, p) $>$ 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 CIGL(a) $\cap$ CIGL(p) $\neq \emptyset$, CIGL(a) $\cap$ CIGL(n) = $\emptyset$, CIGL(p) $\cap$ CIGL(n) = $\emptyset$, and if (cos(E(a), E(p)) $>$ cos(E(a), E(n)) AND levsim(a, p) $>$ levsim(a, n)). Here, 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(RG(a), RG(p)) $>$ J(RG(a), RG(n)) + $\text{margin}_{\text{RG}}$. Here, 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 Huggingface[4]. 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 annotated 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 (Eberts 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(RG(x),RG(y)) ≥ 0.4 or (J(RG(x),RG(y)) ≥ 0.2 and (CXP(x) = CXP(y) or CXB(x) = 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
		R1 (obs)	R1 (anat)	R2	R3	R4	R5	R6	AUC
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 (Alsentzer et al., 2019)	**0.922**	**0.864**	0.933	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 (Boecking 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 ith vector of O and the jth 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.

$$P = \frac{\sum_j \mathbb{1}(\max_i M_{i,j} \geq t)}{|G|} \qquad S_{\text{col}} = \frac{\sum_j \max_i M_{i,j}}{|G|}$$
$$R = \frac{\sum_i \mathbb{1}(\max_j M_{i,j} \geq t)}{|O|} \qquad S_{\text{row}} = \frac{\sum_i \max_j M_{i,j}}{|O|} \tag{607}$$
$$F_1 = 2 \times \frac{P \times R}{P+R} \qquad S = \frac{S_{\text{row}} + S_{\text{col}}}{2}$$

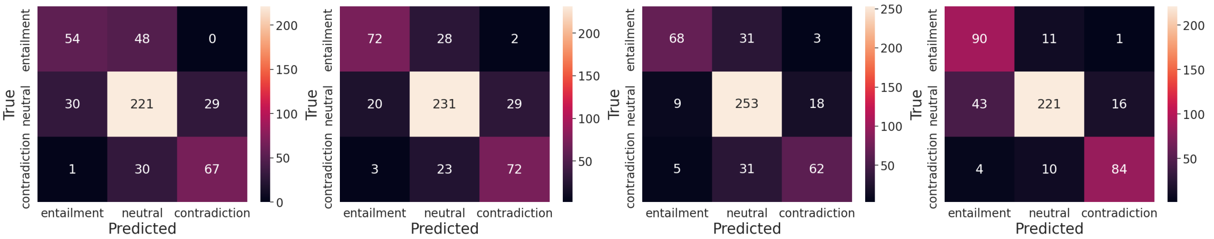
Label extraction results. Table 4 presents results of template-based report generation using various label extraction methods. We report results with the new *CXR Fact Encoder Score* and also include RadGraph metrics (Jaccard similarity, F1 score, Precision, Recall), CheXpert and CheXbert metrics (accuracy, 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. $\mathrm{Acc}^{*}_{\mathrm{E+C}}$ 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	$\mathrm{Acc}_{\mathrm{E}}$	$\mathrm{Acc}_{\mathrm{C}}$	$\mathrm{Acc}_{\mathrm{E+C}}$	$\mathrm{Acc}_{\mathrm{E}}$	$\mathrm{Acc}_{\mathrm{C}}$	$\mathrm{Acc}_{\mathrm{E+C}}$	$\mathrm{Acc}^{*}_{\mathrm{E+C}}$	$\mathrm{Acc}_{\mathrm{E}}$	$\mathrm{Acc}_{\mathrm{C}}$	$\mathrm{Acc}_{\mathrm{E+C}}$	$\mathrm{Acc}^{*}_{\mathrm{E+C}}$
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 (Bannur 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**



CXR Fact Encoder (T+EC+NLI) CXR Fact Encoder (T+C+EC+NLI+SD) CXR Fact Encoder (NLI fine-tuning) GPT-4

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	Jac	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 presented 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 Bressem. 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.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 469–484, Toronto, Canada. Association for Computational Linguistics.

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 *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *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)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2006–2013. IOS Press.

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.

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. *arXiv preprint arXiv:2307.06439*.

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 *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021a. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A. Young, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. 2021b. Visualchexbert: Addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 105–115, New York, NY, USA. Association for Computing Machinery.

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

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

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

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

Fangyu Liu, Qianchu Liu, Shruthi Bannur, Fernando Pérez-García, Naoto Usuyama, Sheng Zhang, Tristan Naumann, Aditya Nori, Hoifung Poon, Javier Alvarez-Valle, Ozan Oktay, and Stephanie L. Hyland. 2023a. Compositional Zero-Shot Domain Transfer with Text-to-Text Models. *Transactions of the Association for Computational Linguistics*, 11:1097–1113.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, YANG FENG, and Zuozhu Liu. 2023b. A chatGPT aided explainable framework for zero-shot medical image diagnosis. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, et al. 2023. Impressiongpt:

an iterative optimizing framework for radiology report summarization with chatgpt. *arXiv preprint arXiv:2304.08448*.

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. *ACM Computing Surveys (CSUR)*, 54(10s):1–40.

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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.

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. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080.

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

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.

Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. 2021. Clinically correct report generation from chest x-rays using templates. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 654–663. Springer.

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

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

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. *arXiv preprint arXiv:2305.03513*.

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. *CoRR*, abs/2004.09167.

Tanveer Syeda-Mahmood, Ken CL Wong, Yaniv Gur, Joy T Wu, Ashutosh Jadhav, Satyananda Kashyap, Alexandros Karargyris, Anup Pillai, Arjun Sharma, Ali Bin Syed, et al. 2020. Chest x-ray report generation through fine-grained label learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 561–571. Springer.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549.

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 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

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

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

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

A Appendix

A.1 CXR Fact Encoder's Tasks Details

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

A second group of tasks are classification tasks (Figure 6). These include category (5 classes: *anatomical finding*, *disease*, *technical assessment*, *tubes and lines* and *device*), health status (4 classes: i.e., *normal*, *abnormal*, *ambiguous*, *unknown*), comparison status (15 classes, see Figure 15), Chest ImaGenome observations (74 classes, see Figure 16) and anatomical locations (38 classes, see Figure 17). Category, Health Status and Comparison Status are single-label multi-class classification tasks, whereas Chest ImaGenome observations and anatomical locations are multi-label multi-class classification tasks. Implementing these tasks require attaching fully connected heads on top of CXR-BERT-specialized's built-in projection layer in order to perform the classification.

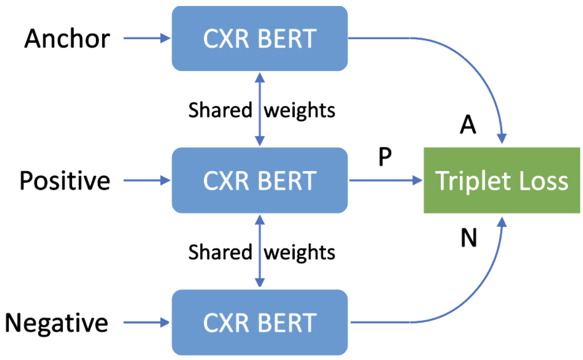


Figure 5: Triplet loss (T)

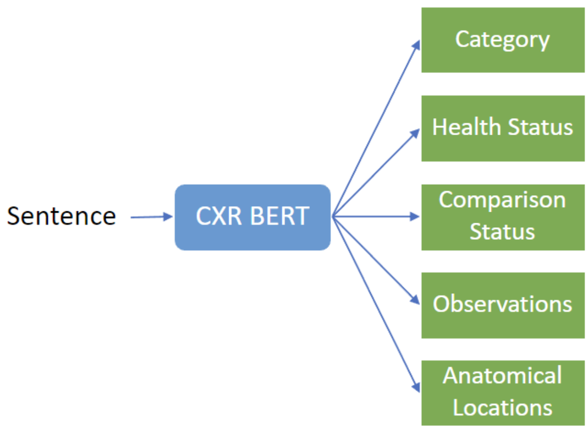


Figure 6: Sentence classification (C)

Another task is sentence decoding (Figure 7). We attach a lightweight, shallow Transformer Decoder to CXR-BERT-specialized's projection layer in order to generate back the original sentence. This can be viewed a sort of text autoenconder, forcing the projection layer to capture as much information as possible of the input sentence to facilitate the reconstruction of the sentence by the Transformer Decoder. We use a Transfomer Decoder with embedding, hidden and feedforward dimension 256, only one self-attention head and only one layer.

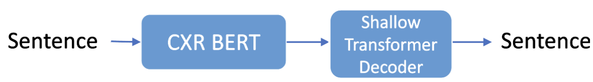


Figure 7: Sentence decoding (SD)

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.

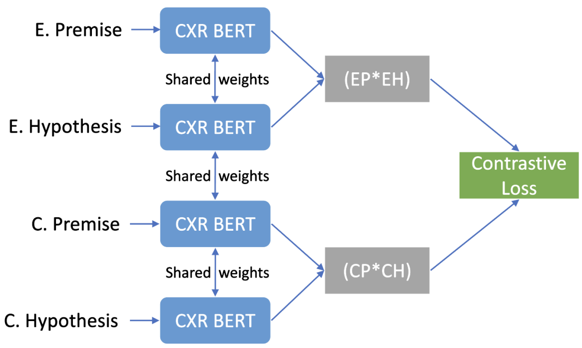


Figure 8: Entailment/contradiction quadruplet loss (EC)

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).

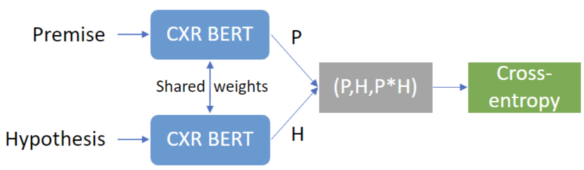


Figure 9: Natural language inference (NLI)

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 clustering[5] over

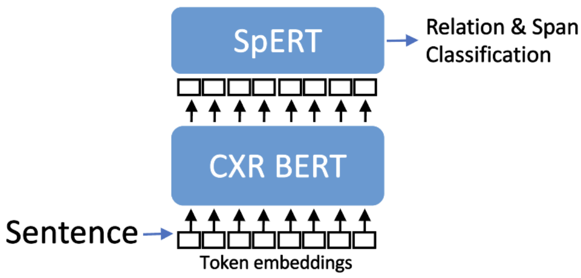


Figure 10: Entity and relation extraction (ER) with SpERT

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

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

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)
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.	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 **CXR Fact Encoder Score (sim):** 1.000 **RadGraph F1:** 0.796 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	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 **CXR Fact Encoder Score (sim):** 0.954 **RadGraph F1:** 0.539 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	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 **CXR Fact Encoder Score (sim):** 0.934 **RadGraph F1:** 0.455 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0
CXR Fact Encoder (f=200, a=50, n=1000)	**CheXbert**	**CheXpert labeler**	**Chest ImaGenome**
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 **CXR Fact Encoder Score (sim):** 0.794 **RadGraph F1:** 0.460 **CheXpert Acc:** 0.929 **CheXbert Acc:** 0.929	the heart is enlarged. the cardiomediastinal 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 **CXR Fact Encoder Score (sim):** 0.517 **RadGraph F1:** 0.021 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	the heart is enlarged. the cardiomediastinal 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 **CXR Fact Encoder Score (sim):** 0.517 **RadGraph F1:** 0.021 **CheXpert Acc:** 1.0 **CheXbert Acc:** 1.0	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 **CXR Fact Encoder Score (sim):** 0.647 **RadGraph F1:** 0.103 **CheXpert Acc:** 0.929 **CheXbert Acc:** 0.929

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

A.4 ChatGPT prompts

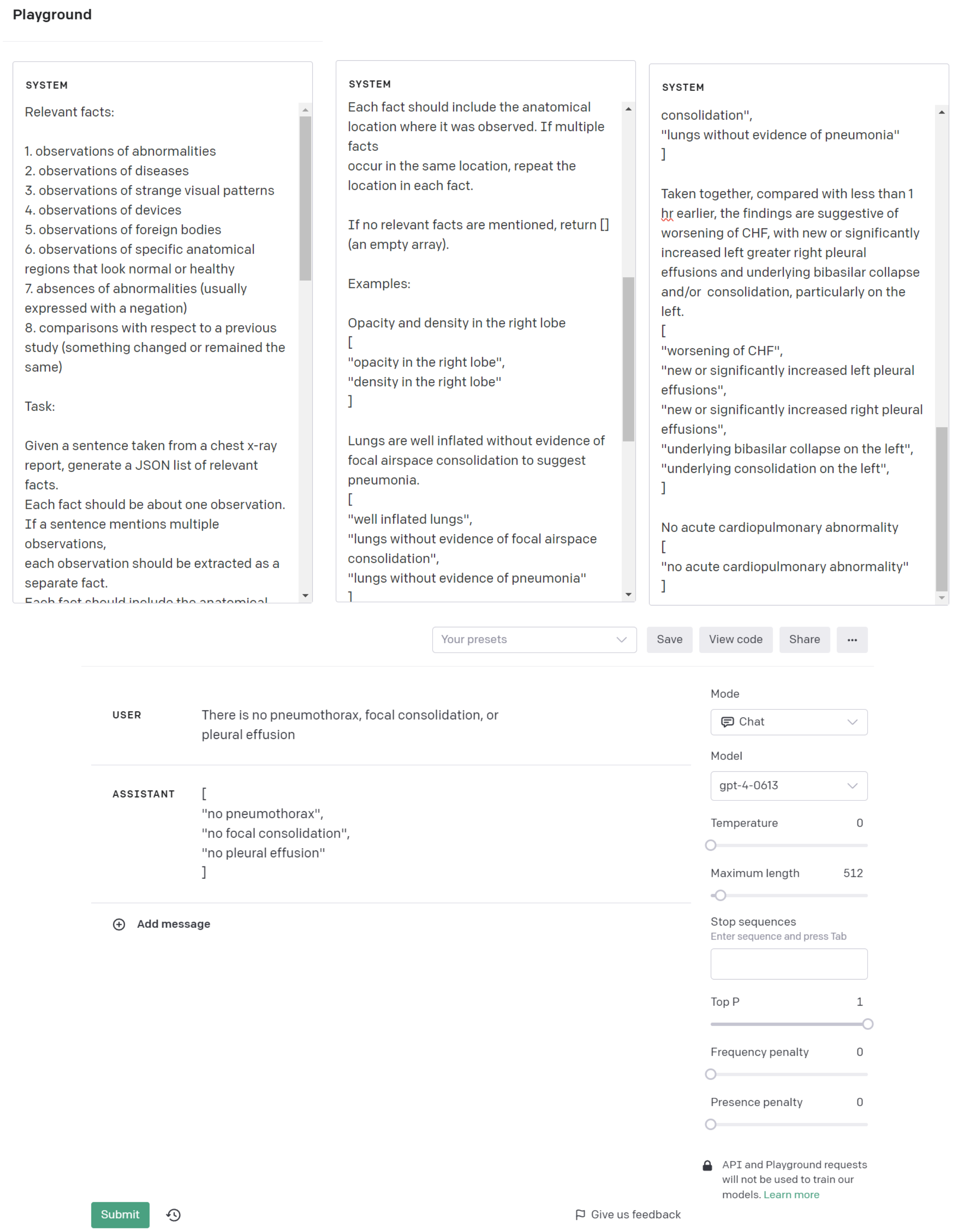


Figure 11: Screenshot of OpenAI's playground web interface showing the prompt used to extract facts from chest X-ray report sentences. The screenshot was manipulated in order to display the full instructions given to ChatGPT.

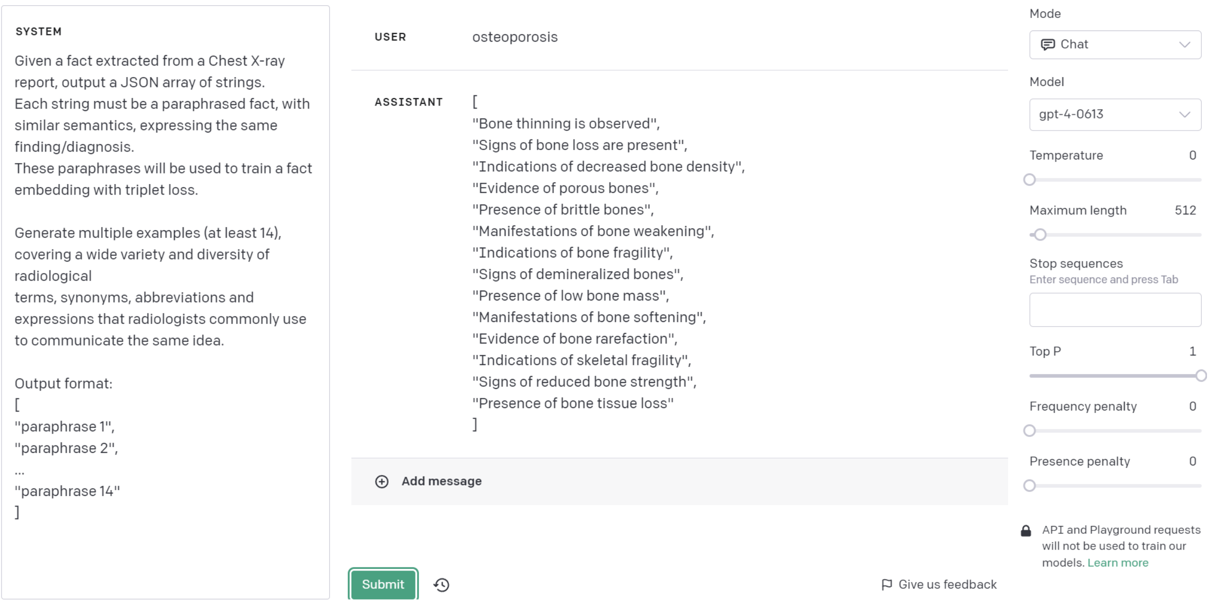


Figure 12: Fact to paraphrases

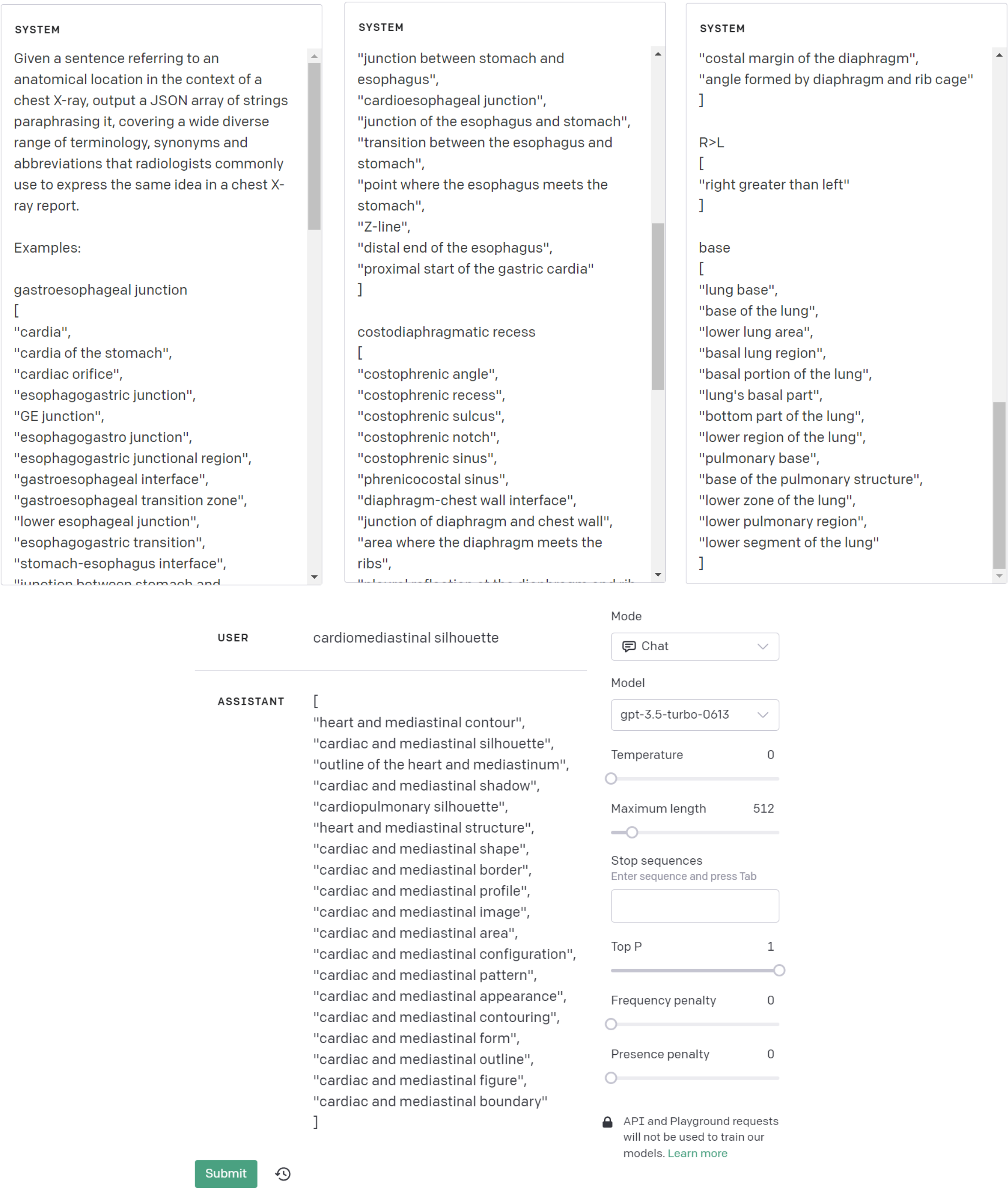


Figure 13: Anatomy location to paraphrases

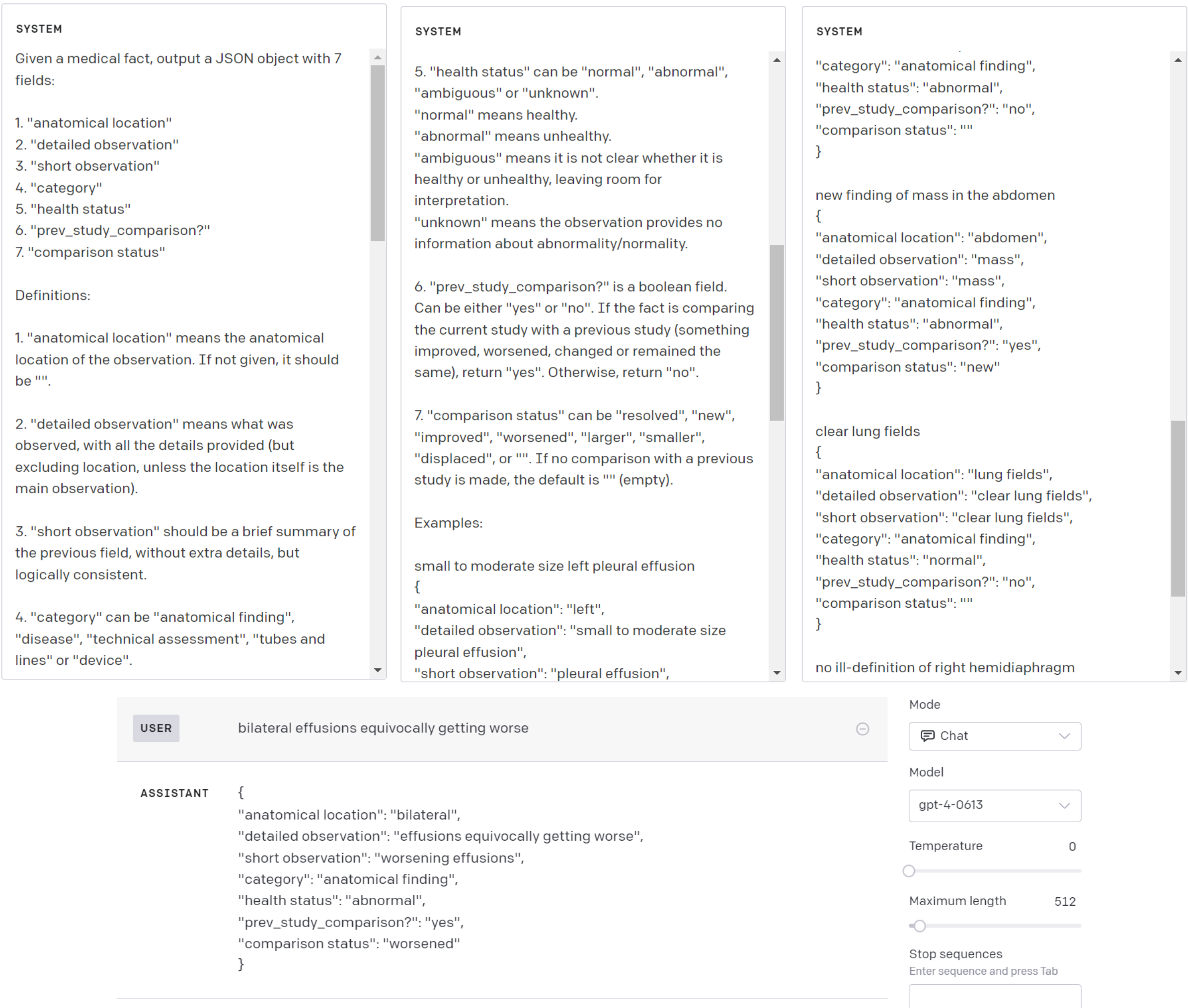


Figure 14: Fact to metadata

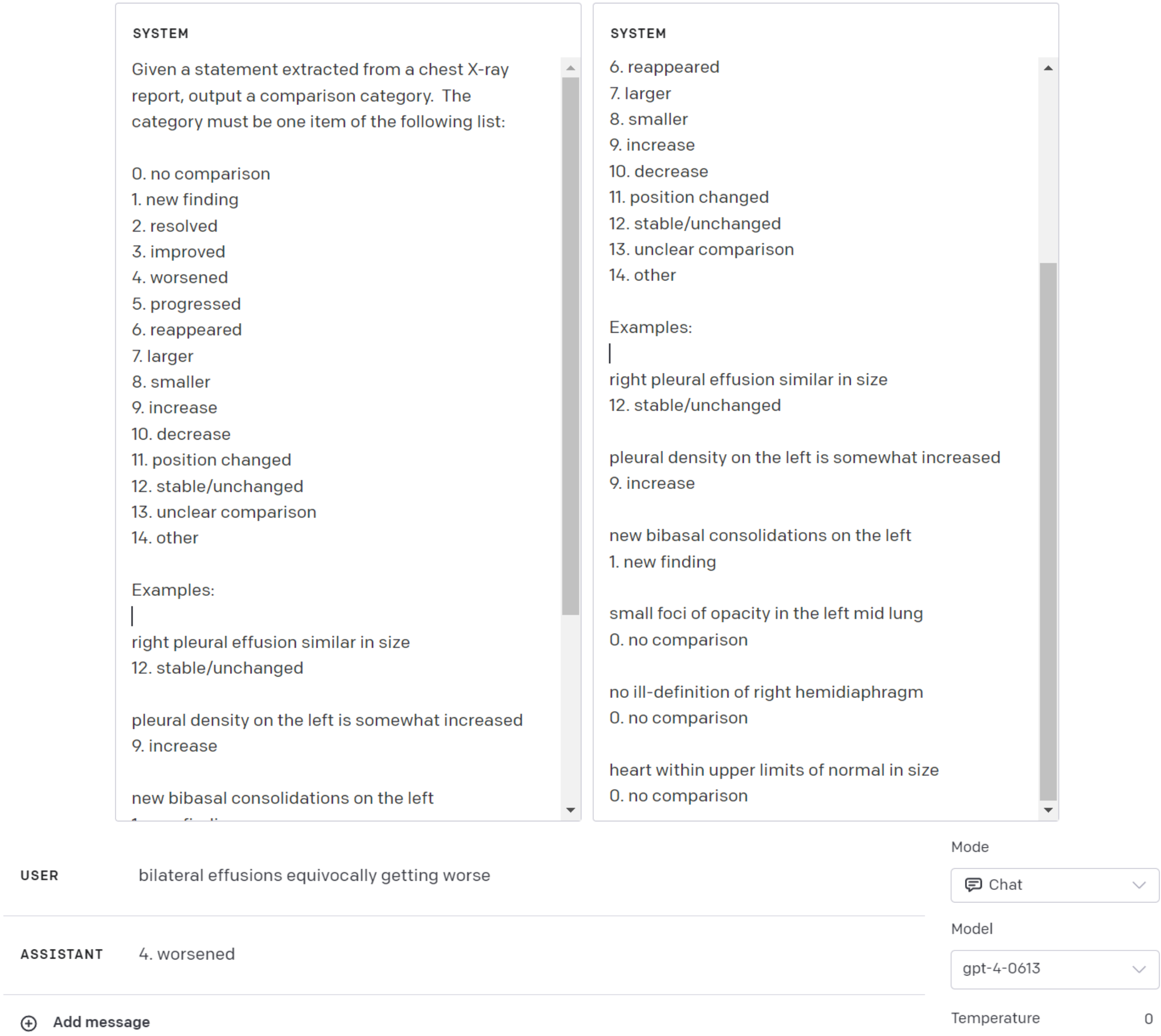


Figure 15: Fact to comparison status

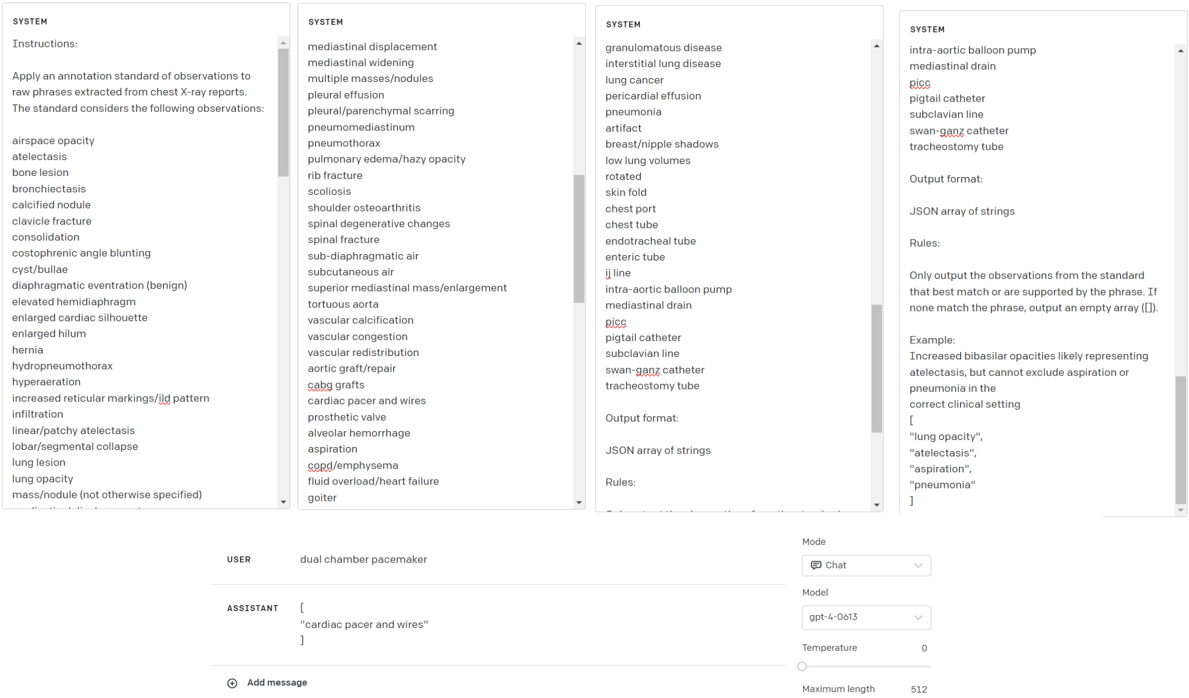


Figure 16: Fact to observations

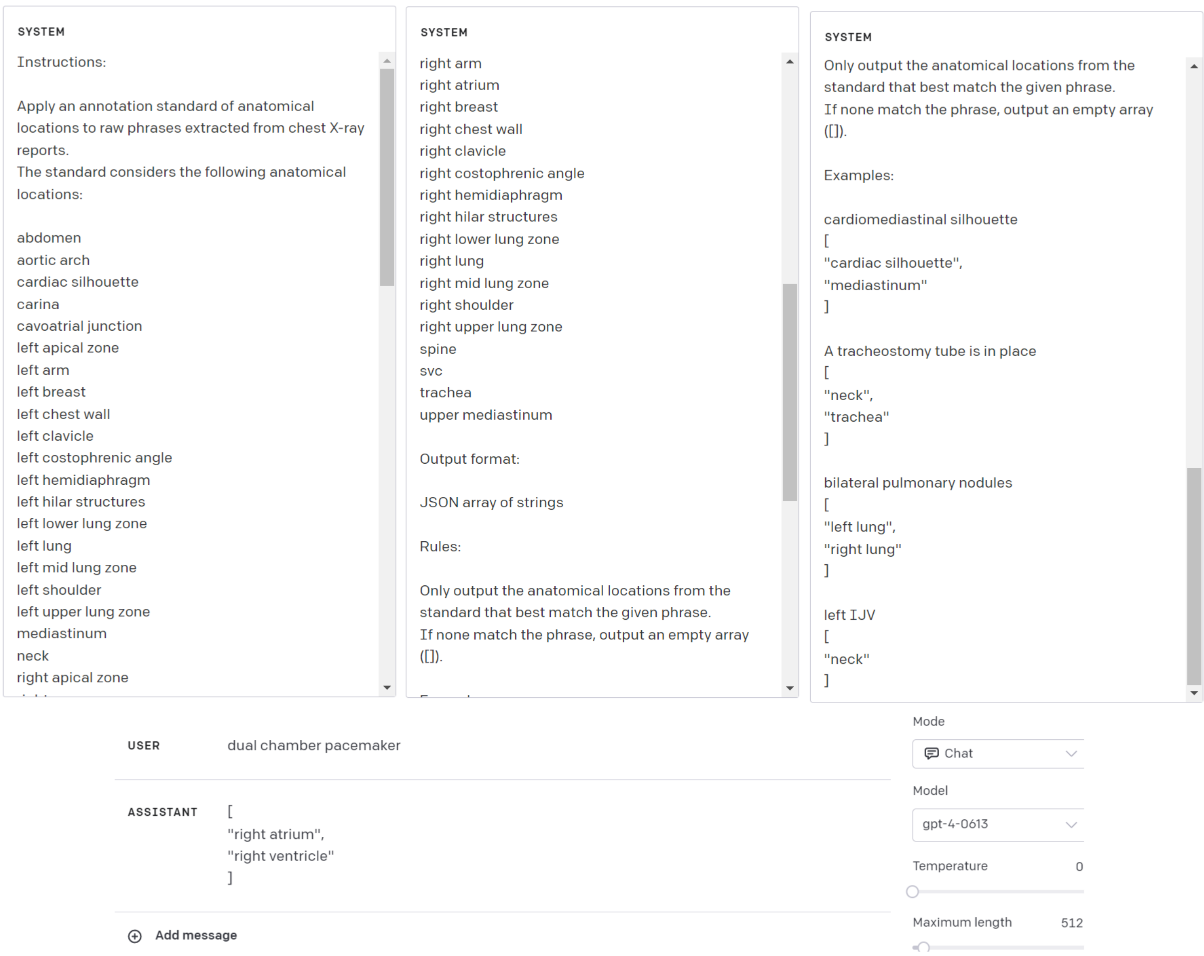


Figure 17: Fact to anatomical locations

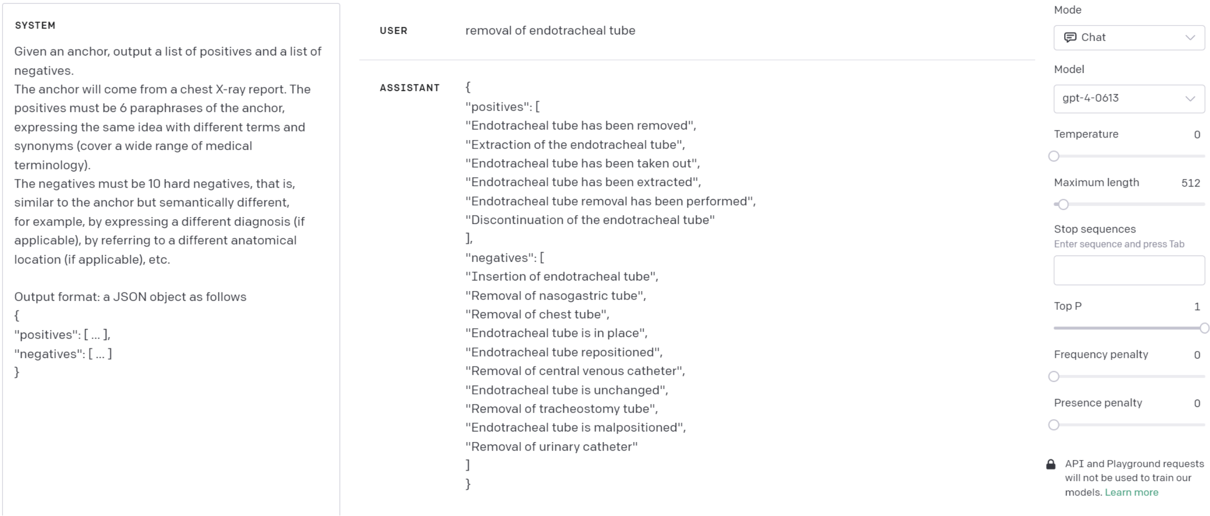


Figure 18: Hard triplets generated by ChatGPT

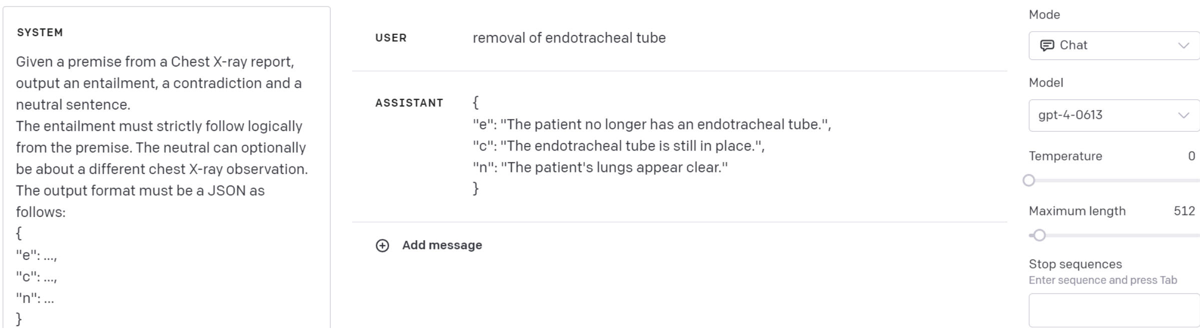


Figure 19: NLI Prompt 1

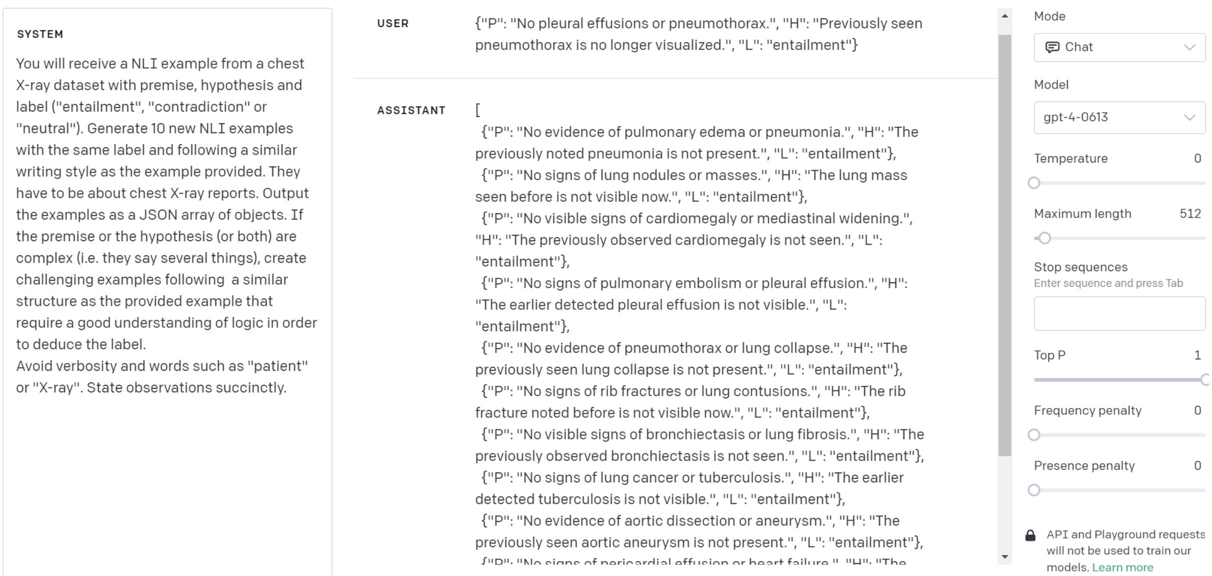


Figure 20: NLI Prompt 2

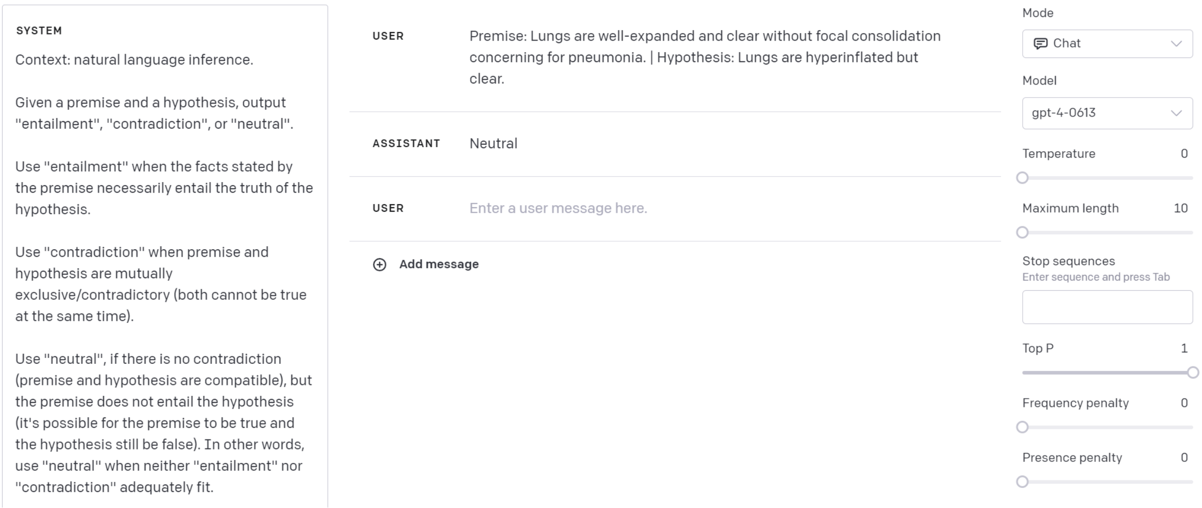


Figure 21: NLI Prompt 3

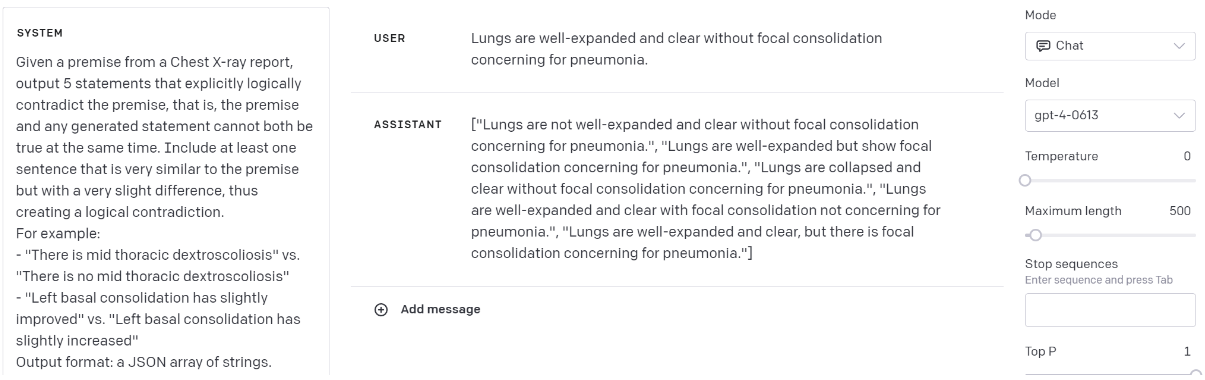


Figure 22: NLI Prompt 4