PromptRad: Knowledge-Enhanced Multi-Label Prompt-Tuning for Low-Resource Radiology Report Labeling

Ying-Jia Lin[1], Tzu-Chin Lo[2]*, Ping-Chien Li[3], Chi-Tung Cheng[4], Chien-Hung Liao[4], Hung-Yu Kao[5]

[1]Department of Artificial Intelligence and AI Research Center, Chang Gung University,
[2]Department of Radiology, Sijhih Cathay General Hospital,
[3]Department of Medical Imaging and Intervention, Chang Gung Memorial Hospital,
[4]Department of Trauma and Emergency Surgery, Chang Gung Memorial Hospital,
[5]Department of Computer Science, National Tsing Hua University
yjlin@cgu.edu.tw, hykao@cs.nthu.edu.tw

Abstract

Automatic report labeling facilitates the identification of clinical findings from unstructured text and enables large-scale annotation for medical imaging research. Existing rule-based labelers struggle with the diverse descriptions in clinical reports, while fine-tuning pre-trained language models (PLMs) requires large amounts of labeled data that are often unavailable in clinical settings. In this paper, we propose PromptRad, a knowledge-enhanced multi-label **prompt**-tuning approach for **rad**iology report labeling under low-resource settings. PromptRad reformulates multi-label classification as masked language modeling and incorporates synonyms from the UMLS Metathesaurus into a multi-word verbalizer to enrich category representations. By fine-tuning the PLM without additional classification layers, PromptRad requires substantially less labeled data than conventional fine-tuning. Experiments on liver CT (computed tomography) reports show that PromptRad outperforms dictionary-based and fine-tuning baselines with only 32 labeled training examples, and achieves competitive performance with GPT-4 despite using a much smaller model. Further analysis demonstrates that PromptRad captures complex negation patterns more effectively than existing methods, making it a promising solution for report labeling in data-scarce clinical scenarios. Our code is available at https://github.com/ila-lab/PromptRad.

1 Introduction

Radiology reports contain valuable medical findings about patients' conditions and play a crucial role in clinical decision-making. However, these reports are usually lengthy and unstructured (see Figure 1 for an example), making it difficult to efficiently extract clinical information (Chen et al., 2018; Pons et al., 2016). Traditional approaches rely on expert knowledge from the Unified Medical Language System (UMLS) (Bodenreider, 2004) and computational linguistics techniques for medical concept extraction (Aronson, 2001; Wang et al., 2017; Peng et al., 2018). However, these dictionary-based methods depend on pre-established mappings between text and medical concepts and often fail when reports use diverse or non-conventional descriptions (Irvin et al., 2019). While pre-trained language models (PLMs) such as BioBERT (Lee et al., 2019) and PubMedBERT (Gu et al., 2021) have been applied to report labeling (Wood et al., 2020; Smit et al., 2020; Li et al., 2022), fine-tuning these models requires substantial labeled data (Devlin et al., 2019; Dodge et al., 2020; Zhang et al., 2021), which is scarce in clinical settings due to the need for domain expertise during annotation.

Example Liver CT Report (de-identified):

CT study of chest and abdomen without and with IV contrast enhancement shows:

1. Rapid enlarging nodule or mass at RUL, suggestive of rapid lung metastasis.
2. Multiple new small nodules (0.2–1.0 cm) at bilateral lungs.
3. A 4cm liver mass at S5, with peripheral globular enhancement, favor hemangioma.
4. Small cysts at right lobe of liver.
5. Gallstones.

Gold Labels: Cyst ✓, Metastasis ✓, Hemangioma ✓

Figure 1: An example liver CT report from our dataset with annotated findings. Underlined terms indicate relevant mentions for each positive label.

Prompt-tuning (Gao et al., 2021; Schick and Schütze, 2021a) offers a promising alternative by transforming classification into masked language modeling, enabling PLMs to perform well with limited labeled data (Schick and Schütze, 2021c). However, existing prompt-tuning methods (Hu et al., 2022; Gao et al., 2021; Liu et al., 2023a) are designed for multi-class classification with mutually exclusive categories. In contrast, radiol-

*Work was done while the author was affiliated with Chang Gung Memorial Hospital.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 235–249

ogy report labeling is inherently a multi-label task, where multiple findings may coexist in a single report. This limits the direct applicability of existing prompt-tuning approaches to report labeling.

We propose PromptRad, a knowledge-enhanced multi-label **prompt**-tuning approach for **rad**iology report labeling under low-resource settings. Here, "low-resource" refers to settings where only a limited number of expert-annotated reports are available for training. PromptRad adapts a pre-trained masked language model (Devlin et al., 2019; Gu et al., 2021) for multi-label classification through prompt-tuning. Inspired by Knowledgeable Prompt-Tuning (KPT) (Hu et al., 2022), we design a multi-word verbalizer that incorporates synonyms from the UMLS Metathesaurus (Bodenreider, 2004) as label-to-word mappings, enriching category representations for clinical contexts. We further develop automatic prompt generation based on Gao et al. (2021) to explore the space of textual templates and enhance performance.

Experiments on liver CT reports from a large medical center demonstrate that PromptRad outperforms dictionary-based methods (Aronson, 2001; Peng et al., 2018) and fine-tuning baselines, and is competitive with GPT-4 (OpenAI, 2023) even with only 32 labeled training examples. Our analysis of negation cases shows that PromptRad captures complex negation patterns more effectively than rule-based approaches, demonstrating its robustness for clinical report labeling.

In summary, our contributions are threefold:

- We formulate low-resource radiology report labeling as a multi-label prompt-tuning problem and adapt a masked language model (Devlin et al., 2019) to predict multiple clinical findings from a single report.
- We introduce a UMLS-informed multi-word verbalizer and an automatic prompt generation strategy to improve label representation and template selection in clinical contexts.
- We evaluate PromptRad on a real-world liver CT report dataset and analyze its behavior under limited supervision, including challenging negation cases.

2 Related Work

2.1 Report Labeling

Traditional approaches to report labeling rely on pre-built knowledge bases. MetaMap (Aronson, 2001) maps medical text to UMLS concepts, and NegBio (Peng et al., 2018) extends it with dependency parsing rules for negation detection. CheXpert (Irvin et al., 2019) replaces predefined concept extractors with manually curated mention lists and more sophisticated negation rules for chest X-ray (CXR) reports. However, these rule-based methods depend on predefined patterns and may fail to generalize across report types and institutions.

Machine learning approaches, including CNN-based (Majkowska et al., 2020; Shin et al., 2017) and LSTM-based methods (Dahl et al., 2021; D'Anniballe et al., 2022), have also been explored for report labeling but require large amounts of labeled data. More recently, PLM-based approaches have shown strong performance: ALARM (Wood et al., 2020) uses BioBERT (Lee et al., 2019) for MRI reports, CheXbert (Smit et al., 2020) uses BlueBERT (Peng et al., 2019) for CXR reports, and other studies have applied BERT (Devlin et al., 2019; Li et al., 2022) and domain-specific PLMs (Nowak et al., 2023) to various report labeling tasks. While effective, fine-tuning PLMs typically requires substantial labeled data (Brown et al., 2020; Dodge et al., 2020; Zhang et al., 2021), which is scarce in clinical settings. Unlike most prior radiology labelers that focus on chest X-ray reports or rely on large annotated corpora, our work focuses on low-resource multi-label labeling for liver CT reports.

2.2 Prompt-Tuning

Prompt-tuning (Schick and Schütze, 2021a; Gao et al., 2021) reformulates classification as masked language modeling, reducing reliance on newly initialized, task-specific classification layers and enabling effective few-shot learning (Schick and Schütze, 2021c). Prompt-based methods have also been explored in biomedical NLP tasks, including biomedical relation extraction (He et al., 2024) and lay summary generation (Wu et al., 2023). KPT (Hu et al., 2022) improves prompt-tuning by incorporating external knowledge into the verbalizer. Wei et al. (2022) extended prompt-tuning to multi-label text classification and proposed PTMLTC (Prompt Tuning Method for Multi-Label Text Classification), which is the closest to our work. However, PTMLTC was designed for the educational domain and relies on a single-word verbalizer, which may not capture the diverse expressions for the target categories. To the best of our knowledge, no prior work has applied prompt-

tuning to multi-label radiology report labeling under low-resource settings, especially using small local models with few-shot training.

2.3 Recent Advances in LLM-based Report Labeling

With the emergence of large language models (LLMs), recent work has explored LLM-based approaches for radiology report analysis. CheXGPT (Gu et al., 2024) uses GPT-4 as a zero-shot labeler and then trains a BERT-based model on 50k pseudo-labeled chest X-ray reports generated by GPT-4. Fytas et al. (2024) propose a multi-turn prompting strategy that combines rule-based insights with LLM predictions for chest X-ray report classification. Domain-adapted models such as Radiology-Llama2 (Liu et al., 2023b) fine-tune open-source LLMs on radiology corpora via instruction tuning, and Abdullah and Kim (2025) fine-tune a smaller LLM for report labeling on the MIMIC-CXR dataset. However, these approaches either require large volumes of (pseudo-)labeled data, domain-specific corpora for continued pretraining, or access to proprietary APIs, raising concerns regarding deployment cost and data privacy in clinical settings. In contrast, PromptRad is a prompt-tuning approach that fine-tunes a masked language model with a UMLS-informed verbalizer, requiring only 32 labeled examples and no access to external APIs.

Table 1: Number of positive findings per category in our labeled clinical report dataset. HCC: Hepatocellular Carcinoma.

Category	Training (2008–2014)	Test (2015–2017)
Cyst	275	109
HCC	214	79
Post-Treatment	205	79
Cirrhosis	168	85
Steatosis	154	93
Metastasis	101	46
Hemangioma	90	29
Total reports	773	325

3 Dataset

This study was approved by the institutional review board of the participating medical center. The dataset contains de-identified reports written in English, covering liver CT examinations from 2008 to 2017 (an example report is shown in Figure 1). The reports were annotated by two senior radiologists for the presence of seven common cancer-related findings. Each finding is labeled as positive (1) or negative (0); suspicious findings were treated as positive to avoid false negatives.

We split the data chronologically to ensure independence between training and test sets. The 2008–2014 subset contains 773 reports and serves as the candidate training pool, while the 2015–2017 subset contains 325 reports and is used as the fixed test set, as summarized in Table 1. To study low-resource scenarios, where only a small number of expert-annotated training examples are available, for each run we randomly sample K reports from the 773-report training pool using stratified sampling (Sechidis et al., 2011) to preserve the class distribution. Our goal is to evaluate how effectively different methods learn under constrained annotation scenarios, rather than to assume that all 773 reports are unavailable.

4 Method

4.1 Problem Definition

Given a radiology report r and a label space $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$ with n distinct classes of findings, the report labeling task is to predict the presence of each finding $y_i \in \mathcal{Y}$ within r. Multiple findings may be present in a single report. We define the training set as $\mathcal{D}_{\text{train}} = \{(r_a, \mathcal{Y}_a)\}_{a=1}^{K}$, where K is the number of training examples and $\mathcal{Y}_a \subseteq \mathcal{Y}$ is the set of findings in report r_a.

4.2 Standard Prompt-Tuning

Prompt-tuning (Schick and Schütze, 2021a; Gao et al., 2021) fine-tunes PLMs without additional classification layers by (1) formatting the input as a cloze test with a textual prompt template containing a `[MASK]` token, and (2) defining a verbalizer that maps each class to a word in the PLM vocabulary. This allows fine-tuning with the MLM objective (Devlin et al., 2019) without introducing new parameters. We illustrate standard prompt-tuning in Figure 2 (a).

4.3 Multi-word Verbalizer of PromptRad

To identify the presence of each finding $y_i \in \mathcal{Y}$ in a report r and transform report labeling into a masked language modeling task (Devlin et al., 2019; Gao et al., 2021; Schick and Schütze, 2021b),

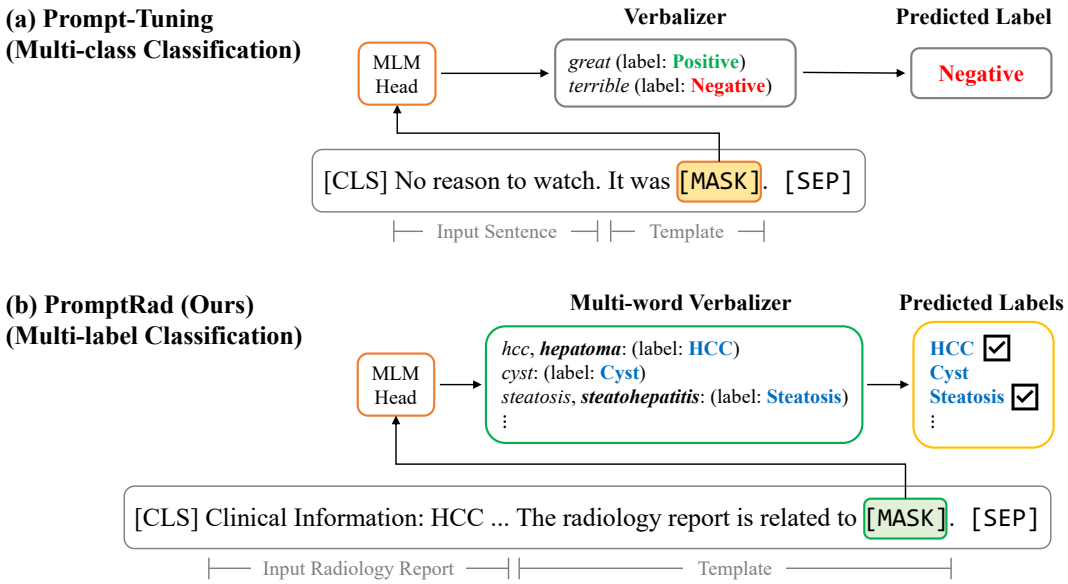

Figure 2: Differences in the approaches of (a) Prompt-tuning for multi-class classification on a general domain task (Gao et al., 2021; Schick and Schütze, 2021b). (b) PromptRad, our method for multi-label classification for radiology report labeling. Bold italic words in the verbalizer denote synonyms augmented from the UMLS Metathesaurus. MLM: masked language model (Devlin et al., 2019).

we define a verbalizer $v : \mathcal{Y} \rightarrow V$ which maps a finding y_i to a word from the vocabulary V of a pre-trained MLM $\mathcal{M}$. We first create a single-word verbalizer, where each finding is mapped to the word that most closely matches its class name (e.g., "Post-treatment" $\rightarrow$ "posttreatment"). Hence, if the model fills the word "posttreatment" in the `[MASK]` position, the report will be classified as containing the "Post-treatment" finding.

In addition to the closest match, synonyms of the class names can also represent the findings. To prevent the model from being biased towards a single mapping per category, we extend the verbalizer with synonyms from SNOMEDCT in the UMLS Metathesaurus (Bodenreider, 2004). For instance, both "hcc" and "hepatoma" serve as mappings for "Hepatocellular Carcinoma," as shown in Figure 2 (b). The complete mappings are listed in Appendix A.1. This multi-word verbalizer can be easily adapted to other report labeling tasks by replacing the mappings with synonyms for the target categories.

4.4 Training

As illustrated in Figure 3, we append a prompt template to each report r to form the input sequence r_p:

r_p = `[CLS]` r The radiology report is related to `[MASK]`. `[SEP]`

where the prompt template is manually designed. Our objective is to determine the presence of each finding by the prediction of the `[MASK]` token in r_p using a pre-trained MLM $\mathcal{M}$.

During training, for each input r_p, we identify the presence of each finding $y_i \in \mathcal{Y}$ by recovering the `[MASK]` token:

$$z_i = f_\mathcal{M}(\texttt{[MASK]}, v(y_i) \mid r_\text{p}) \quad (1)$$

where $f_\mathcal{M}(\texttt{[MASK]}, v(y_i) \mid r_\text{p})$ denotes the logit score of token $v(y_i)$ at the `[MASK]` position produced by the MLM head of $\mathcal{M}$. When multiple label-to-word mappings are available for a finding y_i, the one with the highest likelihood is chosen:

$$z_i = \max_{m=1}^{M_i}(f_\mathcal{M}(\texttt{[MASK]}, v(y_i, m)|r_\text{p})), \quad (2)$$

where M_i is the number of label-to-word mappings for y_i. For example, for "Hepatocellular Carcinoma", M_i is 2 with mappings "hcc" and "hepatoma". Practically, as we show in Figure 3, we create a matrix with the shape of $n \times M$ to facilitate the max operation in Equation 2, where M is the maximum number of mappings among the categories. We pad categories with fewer than M mappings using dummy entries and apply a binary mask so that padded positions do not affect the maximum operation. Then, we optimize $\mathcal{M}$ with the binary cross-entropy loss:

$$\hat{y}_i = \text{sigmoid}(z_i) \quad (3)$$

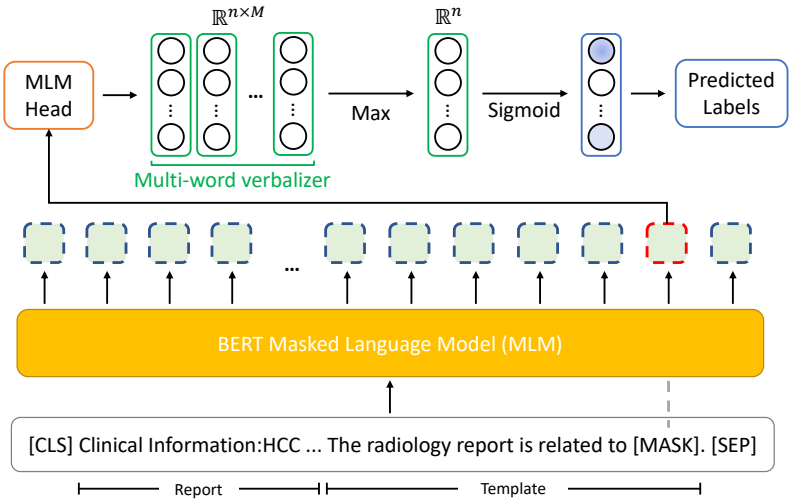

Figure 3: The workflow of the PromptRad report labeling system.

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]. \tag{4}$$

During inference, we assign a finding y_i to a radiology report r if the probability $\hat{y}_i$ exceeds a defined threshold τ for all the categories.

4.5 Automatic Prompt Generation

We argue that better performance can be achieved by training our model with different templates. Instead of relying solely on manually designed templates, we explore automatic prompt generation following Gao et al. (2021).

We take advantage of the generative capability of the pre-trained T5 (Raffel et al., 2020) to automatically create various prompt templates and find the best one for our task. Given a report r, the single-word verbalizer v, and a label y_i, we explore two formats: (1) report first: $[r]$ $[\mathbf{P}_1]$ $[v(y_i)]$ $[\mathbf{P}_2]$, and (2) answer first: $[\mathbf{P}_1]$ $[v(y_i)]$ $[\mathbf{P}_2]$ $[r]$, where $\mathbf{P}_1$ and $\mathbf{P}_2$ are T5-generated prompts placed before and after the label word. Format 1 mirrors our manual template by placing the report before the label. In addition, since the position of the `[MASK]` token can vary in a sequence during the pre-training of BERT (Devlin et al., 2019), we also explore Format 2, where the order is reversed.

For each report with multiple labels, we duplicate the input for each label and use T5 with beam search to generate 100 candidate templates (50 per format), ranked by the joint log-likelihood over all training examples (Gao et al., 2021). We train PromptRad with each candidate and select the best-performing template (Table 6 in Appendix A.2).

5 Experiments

We evaluate PromptRad against a range of baselines on the test set. Unless otherwise specified, all training-based methods (e.g., PubMedBERT and PromptRad) use 32 labeled reports (K=32) as training data. Detailed experimental settings, baseline descriptions, and GPT-4 prompts are provided in Appendix B.

5.1 Performance Comparison

Table 2 shows that the proposed PromptRad method outperforms all the baselines for most of the categories using the manual template, except for GPT-4 (OpenAI, 2023). In addition, Table 2 also shows that PromptRad+AutoT benefits from the proposed automatic template generation method and slightly outperforms GPT-4 in terms of average scores. This result suggests that fine-tuning BERT using appropriate prompts provides a competitive and lightweight alternative to API-based large language models for low-resource report labeling.

For the category of "Hemangioma", "Label Match" can almost perfectly identify the positive findings, showing that the descriptions for "Hemangioma" can be easily identified. The reason that our approaches perform worse than "Label Match" in the "Hemangioma" category is that the number of training examples for "Hemangioma" is relatively small[1], compared to the other categories. We will

[1]During pre-processing, we randomly sampled 32 examples while keeping the distribution of the sampled data the same as that of the original training set, as we mentioned in Section 3. Thus, there are only 2-3 samples labeled as "Hemangioma" each time.

Table 2: Performance comparison in F1-score (%) using the independent test set.

	Cyst	HCC	Post-T	Cirr.	Stea.	Meta.	Hem.	Mac. F1	Mic. F1
GPT-4	86.1	91.5	79.1	96.6	98.9	73.8	95.1	88.7	88.7
GPT-4 (ICL[a])	89.7	88.8	76.3	92.5	92.0	78.5	94.9	87.5	87.4
Label Match	67.5	88.5	0.0	87.7	0.0	48.6	**98.3**	55.8	62.3
MetaMap	77.6	86.9	48.6	53.5	95.7	27.5	84.6	67.8	69.1
NegBio	77.6	86.9	**81.4**	82.7	95.7	27.5	84.6	76.6	79.2
PMB[b]	$53.7_{8.4}$	$80.4_{5.2}$	$70.4_{4.3}$	$64.5_{12.7}$	$48.3_{12.9}$	$54.9_{18.1}$	$37.7_{21.7}$	$58.6_{10.0}$	$60.9_{8.2}$
PMB[b]+MM	$68.1_{5.4}$	$83.9_{1.7}$	$64.5_{2.3}$	$68.0_{6.6}$	$84.5_{6.7}$	$47.1_{10.0}$	$70.1_{15.2}$	$69.5_{5.0}$	$70.3_{3.8}$
PMB[b]+NB	$68.1_{5.4}$	$83.9_{1.7}$	$76.8_{0.3}$	$81.0_{3.7}$	$84.5_{6.7}$	$47.1_{10.0}$	$70.1_{15.2}$	$73.1_{4.5}$	$74.1_{3.5}$
PR[b]	$78.0_{5.3}$	$89.1_{0.8}$	$76.9_{1.8}$	$86.2_{4.2}$	$95.7_{3.7}$	$71.9_{3.5}$	$88.4_{5.6}$	$83.7_{2.1}$	$84.1_{1.7}$
PR+AutoT[b]	$\mathbf{89.5}_{2.4}$	$\mathbf{90.8}_{1.0}$	$78.4_{3.9}$	$\mathbf{91.0}_{3.4}$	$\mathbf{97.3}_{0.9}$	$\mathbf{84.7}_{1.2}$	$92.4_{2.5}$	$\mathbf{89.2}_{1.0}$	$\mathbf{89.4}_{1.0}$

Abbreviations: Cirr.: Cirrhosis; Stea.: Steatosis; Meta.: Metastasis; Hem.: Hemangioma; Mac.: Macro average; Mic.: Micro average; Post-T: Post-Treatment; PMB: PubMedBERT; MM: MetaMap; NB: NegBio; PR: **P**rompt**R**ad (ours).
[a] In-Context Learning with three random samples from the training set. [b] Average from five runs; subscripts: standard deviation.

show the performance of PromptRad with more training examples in Section 5.4. We note that "Label Match" scores zero on "Post-Treatment" and "Steatosis" because these categories are expressed through synonyms (e.g., RFA, TACE, fatty liver) that do not match the category names directly.

5.2 Case Study

To confirm that the proposed method is capable of handling diverse descriptions for the target labels, we perform a case study and list the examples with the corresponding labels in Table 3. In this experiment, we compare the performance of our method with NegBio (Peng et al., 2018) as representative of the dictionary-based methods, since NegBio outperforms MetaMap (Aronson, 2001) in our experiment (Table 2). Such cases are challenging for dictionary-based systems because the target label names may appear in the text even when the corresponding findings are negated or absent.

In Table 3, we observe that most of the positive findings can be correctly identified by both approaches when the label names are explicitly mentioned in the reports. However, NegBio usually fails to negate the findings when the label names show up in reports with negative descriptions, such as "No ... nor HCC in both lobes liver" and "No imaging evidence of cirrhosis ..." in the first two examples of Table 3. Additionally, NegBio can be misled by common negation abbreviations in radiology reports, such as "R/O" (rule out) in the third report of Table 3, while learning-based methods, such as PromptRad, can learn to recognize these abbreviations from the training data.

This experiment demonstrates the potential of the proposed method in handling negation cases. To further validate the effectiveness of the proposed method for handling negation cases in radiology reports, we next perform a quantitative experiment.

Table 3: Case study comparing PromptRad+AutoT with NegBio on reports containing negated findings. Bold italic labels indicate gold annotations for each sentence.

	Segments (*Gold Label*) and Predicted Positive Labels
Report 1	A tiny hepatic cyst in S7/8> Patency of the SMV **(*Positive: Cyst*)** No obvious hypervascular nodule nor HCC in both lobes liver. **(*Negative: HCC*)** Mild liver cirrhosis and portal hypertension. **(*Positive: Cirrhosis*)** **NegBio**: Cyst, HCC, Cirrhosis **PromptRad+AutoT**: Cyst, Cirrhosis
Report 2	No imaging evidence of cirrhosis of liver. **(*Negative: Cirrhosis*)** Two small 0.6-cm and 1.4-cm densely packed lipiodol puddles in S7 without identifiable viable tumor, suggestive of good response to previous TACE without viability. **(*Positive: HCC, Post-Treatment*)** Multiple hepatic cysts are noted. **(*Positive: Cyst*)** **NegBio**: Cyst, HCC, Post-Treatment, Cirrhosis **PromptRad+AutoT**: Cyst, HCC, Post-Treatment
Report 3	Remarkable fatty liver with GB stones; **(*Positive: Steatosis*)** R/O metastasis in the anterior abdominal wall. **(*Negative: Metastasis*)** **NegBio**: Steatosis, Metastasis **PromptRad+AutoT**: Steatosis

5.3 Effectiveness of Negation Handling

To quantitatively assess the effectiveness of our proposed method in handling negation cases in radiology reports, we collect the reports from the test set that explicitly mentioned the label names, but were annotated as negative findings by our radiologists. We excluded the “Cyst", “Post-Treatment", “Steatosis", and “Hemangioma" categories, as reports mentioning these label names usually indicate positive findings of the corresponding categories.

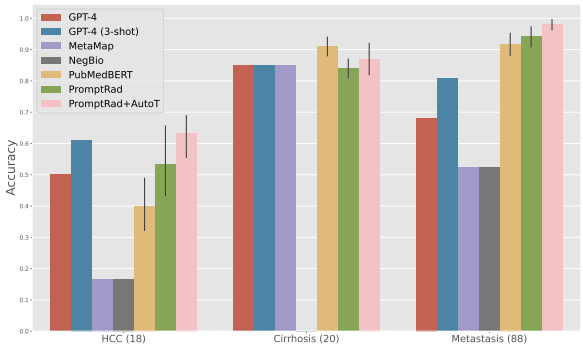

Figure 4: Accuracy on negation cases: reports that mention a finding name but are annotated as negative. Values in parentheses indicate the number of cases per category.

Figure 4 shows that PromptRad+AutoT (ours) is better at handling negation cases than GPT-4 (OpenAI, 2023), NegBio (Peng et al., 2018), and MetaMap (Aronson, 2001) for the negated reports in the three categories. Additionally, our manual-template approach (PromptRad) remains competitive compared to the other baselines. To our surprise, the performance of NegBio on “Cirrhosis" is unsatisfactory. This is due to the fact that NegBio relies heavily on natural language rules of dependency parsing and takes inputs with complete sentence structures. However, the descriptions relevant to “Cirrhosis" in our dataset are usually short (e.g., *No liver cirrhosis.*), which makes it difficult for NegBio to identify the negation cases. Besides, though our methods are slightly inferior to PubMedBERT (Gu et al., 2021) for the reports with “Cirrhosis", they are still much better at handling negation cases than PubMedBERT for the reports with “HCC" and “Metastasis". Based on the overall performance in Figure 4, we can still conclude that our method is a better choice for handling negation cases in radiology reports than the other baselines.

5.4 Performance Comparison for Different Training Sizes

In the previous sections, we studied the performance of the proposed methods for using only 32 labeled reports for training. In this section, we investigate the impact of the training size on the performance of the report labeling task. We randomly sampled different numbers of reports from the training set and evaluated the performance of PromptRad and PubMedBERT on the test set. Figure 5 shows that both methods exhibit a similar trend in performance improvement with the increase of training sizes, and both methods achieve more than 90% F1-score in macro average when the training size is larger than 128 samples. In addition, the proposed method outperforms PubMedBERT in almost all categories with training sizes smaller than 128 samples, showing that our method is effective under data scarcity scenarios.

5.5 Ablation Study for the Verbalizer

Since PromptRad and PromptRad+AutoT use the multi-word verbalizer, to confirm the effectiveness of the multi-word verbalizer, we provide a performance comparison for the two types of verbalizers in Table 4. We find that our methods using the multi-word verbalizer outperform the ones using the single-word verbalizer on average across all seven categories. Notably, the multi-word verbalizer also reduces variance across runs (e.g., Cirrhosis std. drops from 12.6 to 4.2 for PromptRad), suggesting that multiple mappings provide more stable category representations during few-shot training. Therefore, we conclude that the inclusion of multiple label-to-word mappings is beneficial for the radiology report labeling task.

We further note that, according to the complete mappings in Appendix A.1, the two verbalizers differ only for “Hepatocellular Carcinoma” and “Steatosis”. Since PromptRad is trained with shared PubMedBERT parameters and a shared MLM head, these mapping changes may indirectly affect other categories through the joint multi-label objective. Thus, differences for categories with identical mappings likely reflect shared-parameter effects. For other report labeling tasks, the multi-word verbalizer can be easily extended by adding synonyms for additional target categories when available.

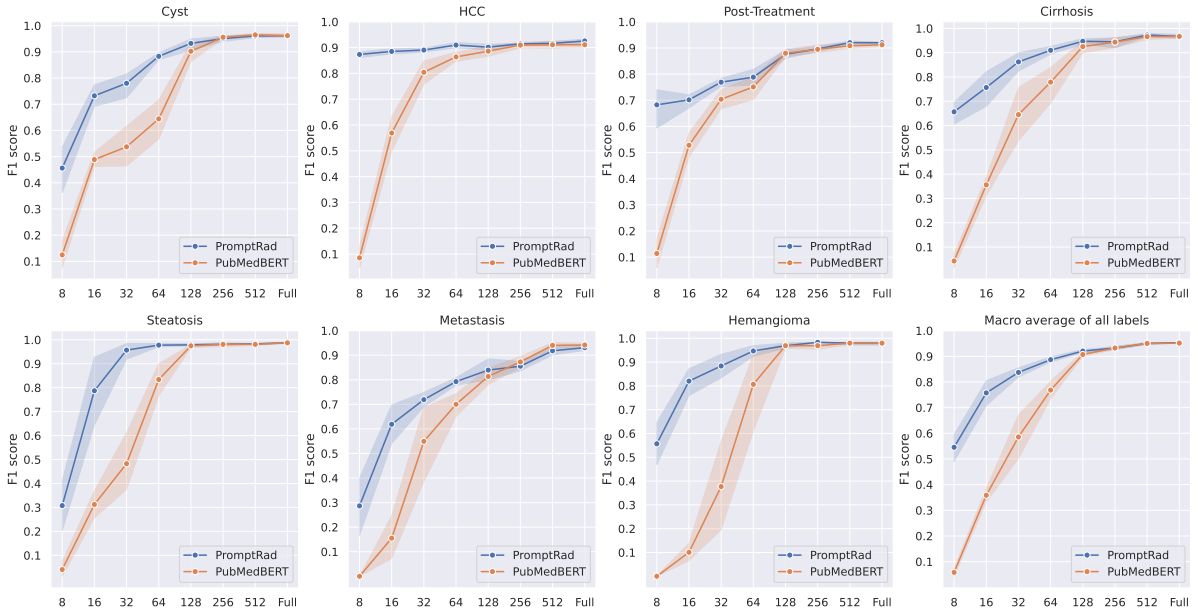

Figure 5: Comparison in F1-score for using different numbers of reports for training. We ran experiments five times for each training size, and for a training size of 8, we repeated the experiment ten times. "Full" indicates the use of the full training set.

Table 4: Performance comparison in F1-score for PromptRad and PromptRad+AutoT using the single-word and multi-word verbalizers.

	PromptRad		PromptRad+AutoT	
Labels	Single	Multi	Single	Multi
Cyst	$79.1_{4.2}$	$78.0_{5.3}$	$89.4_{3.2}$	$\mathbf{89.5}_{2.4}$
HCC	$89.4_{0.6}$	$89.1_{0.8}$	$89.9_{0.6}$	$\mathbf{90.8}_{1.0}$
Post-Treatment	$74.4_{2.4}$	$\mathbf{76.9}_{1.8}$	$\mathbf{80.2}_{3.7}$	$78.4_{3.9}$
Cirrhosis	$84.2_{12.6}$	$\mathbf{86.2}_{4.2}$	$90.3_{2.5}$	$\mathbf{91.0}_{3.4}$
Steatosis	$96.3_{2.2}$	$95.7_{3.7}$	$95.7_{2.2}$	$\mathbf{97.3}_{0.9}$
Metastasis	$74.0_{5.8}$	$71.9_{3.5}$	$83.2_{1.7}$	$\mathbf{84.7}_{1.2}$
Hemangioma	$83.6_{5.8}$	$\mathbf{88.4}_{5.6}$	$89.1_{7.8}$	$\mathbf{92.4}_{2.5}$
Macro Avg.	$83.0_{3.0}$	$\mathbf{83.7}_{2.1}$	$88.2_{1.6}$	$\mathbf{89.2}_{1.0}$
Micro Avg.	$84.1_{2.7}$	$\mathbf{84.1}_{1.7}$	$88.9_{1.4}$	$\mathbf{89.4}_{1.0}$

Scores were averaged from five runs. Subscripts: standard deviation.

6 Discussion

6.1 Why PromptRad Works

PromptRad outperforms standard BERT fine-tuning with PubMedBERT (Gu et al., 2021) across all categories when only 32 labeled reports are available (Table 2), and this advantage persists for training sizes below 128 (Figure 5). We attribute this to the alignment between prompt-tuning and masked language model pre-training. Unlike standard fine-tuning, which introduces a randomly initialized task-specific classification layer, PromptRad introduces no additional task-specific classification parameters. It reuses the pre-trained MLM head and predicts labels through verbalized clinical concepts, preserving closer alignment with the masked language modeling objective used during pre-training.

6.2 Advantages over Rule-based and API-based Alternatives

Compared to dictionary-based systems such as MetaMap (Aronson, 2001) and NegBio (Peng et al., 2018), PromptRad learns from a small number of labeled reports and better handles diverse clinical descriptions and negation patterns (Figure 4), which rule-based systems often miss due to reliance on predefined patterns.

PromptRad also achieves competitive performance with GPT-4 (OpenAI, 2023) under our low-resource setting while using a lightweight PubMedBERT backbone. Beyond efficiency, a locally deployable model avoids sending sensitive clinical reports to external APIs, which is an important consideration for real-world clinical deployment. We therefore view PromptRad not as a replacement for general-purpose large language models, but as a practical alternative when data privacy, deployment cost, and limited annotation resources are central constraints.

6.3 Comparison with Local LLMs

A natural alternative to API-based LLMs is deploying open-source LLMs locally (e.g., Llama 3

(Grattafiori et al., 2024), Qwen (Yang et al., 2025)). While this addresses privacy concerns, it requires substantial GPU memory (typically 16GB+ VRAM even for 7–8B models) that may be unavailable in smaller clinical institutions. In contrast, PromptRad uses a 110M-parameter backbone that can be deployed on consumer-grade hardware, including CPU-only inference. This makes PromptRad practical for resource-constrained clinical environments where both privacy and deployment cost are concerns.

7 Conclusion

In this paper, we introduce PromptRad, a knowledge-enhanced prompt-tuning approach for multi-label radiology report labeling under low-resource settings. Using a UMLS-based multi-word verbalizer and only 32 labeled reports, PromptRad outperforms dictionary-based and fine-tuning baselines, achieves competitive performance with GPT-4, and handles complex negation patterns more effectively than existing methods. As a lightweight, locally deployable model, PromptRad offers a practical solution for clinical report labeling where data privacy and annotation costs are central concerns, and can further support applications such as benign lesion retrieval and large-scale radiograph annotation.

Limitations

This study is based on liver CT reports from a single medical center, which may limit the generalizability of PromptRad to reports from other institutions, imaging domains, or reporting styles. In addition, our experiments focus on English reports and seven predefined liver-related findings, so further validation is needed to assess whether the method transfers well to broader clinical settings. Furthermore, the number of reports included in the negation-focused analysis is relatively small, as indicated by the counts in Figure 4. Larger annotated datasets are needed to better characterize model robustness on challenging negation cases. Finally, the construction of the multi-word verbalizer also depends on an external knowledge base such as the UMLS Metathesaurus. In settings where suitable terminology resources are unavailable, the applicability of this design may be reduced.

Acknowledgments

This work was supported by the National Science and Technology Council in Taiwan, under grants NSTC 114-2223-E-007-011 and NSTC 114-2222-E-182-001-MY2. We would like to express our deepest gratitude to the reviewers for their thoughtful comments and valuable feedback. We would also like to thank Szu-Tung Lin for his assistance in developing the data annotation platform.

References

Abdullah Abdullah and Seong Tae Kim. 2025. Automated radiology report labeling in chest x-ray pathologies: Development and evaluation of a large language model framework. *JMIR Med Inform*, 13:e68618.

AR Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proceedings. AMIA Symposium*, page 17—21.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Matthew C. Chen, Robyn L. Ball, Lingyao Yang, Nathaniel Moradzadeh, Brian E. Chapman, David B. Larson, Curtis P. Langlotz, Timothy J. Amrhein, and Matthew P. Lungren. 2018. Deep learning to classify radiology free-text reports. *Radiology*, 286(3):845–852. PMID: 29135365.

Fredrik A Dahl, Taraka Rama, Petter Hurlen, Pål H Brekke, Haldor Husby, Tore Gundersen, Øystein Nytrø, and Lilja Øvrelid. 2021. Neural classification of norwegian radiology reports: using nlp to detect findings in ct-scans of children. *BMC Medical Informatics and Decision Making*, 21:84.

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.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *CoRR*, abs/2002.06305.

Vincent M D'Anniballe, Fakrul Islam Tushar, Khrystyna Faryna, Songyue Han, Maciej A Mazurowski, Geoffrey D Rubin, and Joseph Y Lo. 2022. Multi-label annotation of text reports from computed tomography of the chest, abdomen, and pelvis using deep learning. *BMC Medical Informatics and Decision Making*, 22:102.

Panagiotis Fytas, Anna Breger, Ian Selby, Simon Baker, Shahab Shahipasand, and Anna Korhonen. 2024. Can rule-based insights enhance LLMs for radiology report classification? introducing the RadPrompt methodology. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 212–235, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jawook Gu, Kihyun You, Han-Cheol Cho, Jiho Kim, Eun Kyoung Hong, and Byungseok Roh. 2024. Chex-gpt: Harnessing large language models for enhanced chest x-ray report labeling. *Preprint*, arXiv:2401.11505.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Jianping He, Fang Li, Jianfu Li, Xinyue Hu, Yi Nian, Yang Xiang, Jingqi Wang, Qiang Wei, Yiming Li, Hua Xu, and Cui Tao. 2024. Prompt Tuning in Biomedical Relation Extraction. *Journal of Healthcare Informatics Research*, 8(2):206–224.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597.

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

Jia Li, Yucong Lin, Pengfei Zhao, Wenjuan Liu, Linkun Cai, Jing Sun, Lei Zhao, Zhenghan Yang, Hong Song, Han Lv, and Zhenchang Wang. 2022. Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (bert) and in-domain pre-training (idpt). *BMC Medical Informatics and Decision Making*, 22:200.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, and 2 others. 2023b. Radiology-llama2: Best-in-class large language model for radiology. *Preprint*, arXiv:2309.06419.

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

Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, and 1 others. 2020. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431.

S Nowak, David Biesner, YC Layer, M Theis, Helen Schneider, W Block, Benjamin Wulff, UI Attenberger, Rafet Sifa, and AM Sprinkart. 2023. Transformer-based structuring of free-text radiology report databases. *European Radiology*, 33(6):4228–4236.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits on Translational Science proceedings*, 2017:188—196.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *Preprint*, arXiv:2106.03598.

Ewoud Pons, Loes M. M. Braun, M. G. Myriam Hunink, and Jan A. Kors. 2016. Natural language processing in radiology: A systematic review. *Radiology*, 279(2):329–343. PMID: 27089187.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021c. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bonggun Shin, Falgun H. Chokshi, Timothy Lee, and Jinho D. Choi. 2017. Classification of radiology reports using neural attention models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4363–4370.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, Los Alamitos, CA, USA. IEEE Computer Society.

Liting Wei, Yun Li, Yi Zhu, Bin Li, and Lejun Zhang. 2022. Prompt tuning for multi-label text classification: How to link exercises to knowledge concepts? *Applied Sciences*, 12(20).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

David A. Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, Martin Kiik, Keena Patel, Gareth Barker, Sebastian Ourselin, James H. Cole, and Thomas C. Booth. 2020. Automated labelling using an attention model for radiology reports of mri scans (alarm). In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 811–826. PMLR.

Yu-Hsuan Wu, Ying-Jia Lin, and Hung-Yu Kao. 2023. IKM_Lab at BioLaySumm task 1: Longformer-based prompt tuning for biomedical lay summary generation. In *Proceedings of the 22nd Workshop*

on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 602–610, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*.

A Appendix

A.1 Verbalizer Mappings

As shown in Table 5, the differences between the single-word and multi-word verbalizers fall into the categories of "Hepatocellular Carcinoma" and "Steatosis", as there are no additional synonyms for the remaining categories.

A.2 Automatically Generated Templates

We list the automatically generated templates with the highest scores in Table 6. We observe that the T5 model (Phan et al., 2021) tends to generate short but concise prompts, such as "Hepatic `[MASK]`" or "Liver `[MASK]`", while the manual template (also listed in Table 6) is more descriptive.

B Implementation Details

We implement our model using the HuggingFace Transformers (Wolf et al., 2020) (version: 4.12.2) and the PyTorch (Paszke et al., 2019) (version: 1.10.0) libraries. We used NVIDIA GeForce GTX 1070 and RTX 2080 Ti GPUs for training our model and finding templates with automatic prompt generation under the CUDA version of 10.1.

To ensure fair comparisons, all experiments for PromptRad and PubMedBERT use the model checkpoint of the pre-trained PubMedBERT-base[2] (Gu et al., 2021) for initialization, and we use SciFive-base[3] (Phan et al., 2021) for automatic prompt generation. For optimizing the models, we use AdamW (Loshchilov and Hutter, 2019) as the optimizer to train the models. All experiments are reported with 32 labeled reports as training data, unless otherwise specified. To simulate the real-world scenario of only limited labeled data available, we do not use a development set for hyperparameter tuning. Instead, we choose the best configuration according to the training loss. We use grid search with the following ranges:

- Learning rate: $[2e-5, 3e-5, 5e-5]$
- Batch size: $[2, 4, 8]$
- Threshold τ: $[0.2, 0.3, 0.4, 0.5]$
- Ratio of training steps for learning rate warmup: $[0.0, 0.1]$

B.1 Baseline Methods

This section describes the baseline methods we used for comparison in our experiments.

B.1.1 GPT-4

We used the OpenAI API[4] to test GPT-4 on report labeling, with the model name[5] of "gpt-4". Our input instructions can be found in Table 7, where they are displayed as both the system message and the user message. For each liver CT report, we first replaced **[REPORT]** in the user message with the actual report text. The modified user message was then processed by GPT-4 to produce the corresponding output assistant message.

B.1.2 GPT-4 (ICL)

The use of in-context learning (ICL) (Brown et al., 2020; OpenAI, 2023) helps large language models achieve better performance by providing a few labeled examples in an input prompt. In this study, we also compare our approach with GPT-4 (ICL). We provide three random examples (3-shot) from the training set in the user message (Table 7) to the GPT-4 model to see if the performance can be improved through in-context learning.

B.1.3 Label Match

We create a straightforward rule-based baseline that identifies positive findings when a report contains the corresponding label names. Our matching rules also account for plural forms of label names. For instance, an occurrence of "metastases" in a report is also recognized as a positive finding for "metastasis". This simple baseline can also be used to observe the difficulties of each category.

[2] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract

[3] https://huggingface.co/razent/SciFive-base-Pubmed_PMC

[4] https://openai.com/blog/openai-api

[5] https://platform.openai.com/docs/models

Table 5: The label-to-word mappings (verbalizer) we used for PromptRad and PromptRad+AutoT.

Category	Single-Word Verbalizer	Multi-Word Verbalizer
Cyst	cyst	cyst
Hepatocellular Carcinoma	hcc	hcc, **hepatoma**
Cirrhosis	cirrhosis	cirrhosis
Post-Treatment	posttreatment	posttreatment
Steatosis	steatosis	steatosis, **steatohepatitis**
Metastasis	metastasis	metastasis
Hemangioma	hemangioma	hemangioma

Table 6: The top-5 templates from automatic template generation along with the manual template for comparison. "Manual" was used in PromptRad, while "AutoT[1]" was used in PromptRad+AutoT.

Template	Format
Manual	[*Report*] The liver radiology report is related to `[MASK]`.
AutoT[1]	Hepatic `[MASK]`: [*Report*]
AutoT[2]	Liver `[MASK]`: [*Report*]
AutoT[3]	[*Report*] Hepatic `[MASK]`.
AutoT[4]	Abdominal `[MASK]`. [*Report*]
AutoT[5]	Liver `[MASK]` in [*Report*]

B.1.4 MetaMap

MetaMap (Aronson, 2001) is a widely used dictionary-based tool for extracting concepts from biomedical text by mapping text to the UMLS Metathesaurus (Bodenreider, 2004). Due to its vast repository of biomedical knowledge and its integration of both hierarchical and non-hierarchical relationships using semantic types and relations, MetaMap stands as a fundamental and effective benchmark for report labeling tasks (Wang et al., 2017). We use the 2020 version of MetaMap with pymetamap[6], an open source MetaMap wrapper, to extract the concepts for each report in our test set.

For each report, we obtain a positive finding if MetaMap returns one of the corresponding UMLS CUIs (Concept Unique Identifiers) of each category. The mappings between CUIs and categories that we used for MetaMap to retrieve the positive findings in our task are provided in Table 8.

B.1.5 NegBio

NegBio (Peng et al., 2018) is a dictionary-based tool for negation and uncertainty detection in clinical text. This approach should work in conjunction with either CheXpert (Irvin et al., 2019) or MetaMap (Aronson, 2001). We use the official implementation of NegBio[7] to extract the concepts for each report in our test set using the same CUIs in Table 8.

B.1.6 PubMedBERT

We fine-tune the pre-trained PubMedBERT model (Gu et al., 2021) for multi-label classification using the HuggingFace API (Wolf et al., 2020). This baseline serves as a representative of the standard fine-tuning approach for PLMs. Similar to the proposed method, we set a threshold τ to determine the positive findings for each label. We perform hyperparameter tuning for PubMedBERT with the same hyperparameter ranges as the proposed method, as described in Section B.

B.1.7 PubMedBERT+MM

We create an ensemble of PubMedBERT (Gu et al., 2021) and MetaMap (MM) (Aronson, 2001) by combining the predicted positive findings from PubMedBERT and MetaMap, denoted "PubMedBERT+MM". This baseline serves as a hybrid model that combines the strengths of both expert knowledge from MetaMap and the machine learning ability of PubMedBERT.

[6] https://github.com/AnthonyMRios/pymetamap

[7] https://github.com/ncbi-nlp/NegBio

B.1.8 PubMedBERT+NB

In addition to PubMedBERT+MM, we also create "PubMedBERT+NB", an ensemble of PubMedBERT (Gu et al., 2021) and NegBio (NB) (Peng et al., 2018). As with PubMedBERT+MM, PubMedBERT+NB combines the predicted positive findings from PubMedBERT and NegBio to take advantage of both methods.

B.2 GPT-4 Prompt Design

Table 7 shows the input message we used to query GPT-4 via the OpenAI API for report labeling. The system message assigns the role of a radiologist, and the user message provides the classification task along with the report text. For GPT-4 (ICL), we additionally include three randomly sampled labeled examples from the training set in the user message.

B.3 UMLS Concept Identifiers

Table 8 lists the UMLS Concept Unique Identifiers (CUIs) used with MetaMap and NegBio to identify positive findings for each category. Since "Post-Treatment" encompasses multiple treatment procedures, we consider a report positive for this category if any of the corresponding CUIs is detected.

B.4 Evaluation Metric

We use the F1-score as the evaluation metric for the seven categories in our experiments. We also include the micro-averaged F1-score and macro-averaged F1-score for the overall performance of each model. All the experiments use the same test set mentioned in Section 3 for evaluations. For the training-based methods, including PubMedBERT (Gu et al., 2021) and our approaches, we report the average scores and standard deviations of 5 runs with different random seeds. For the experiment with only eight training examples, we report the results of 10 runs with different random seeds. This is because the performance of the training-based methods can be more unstable when the training size is smaller, and we want to provide a more comprehensive evaluation for this experiment.

Table 7: Input message and an example output of GPT-4 using the OpenAI API.

Role	**Content**
System	You are a professional radiologist who knows computed tomography (CT) very much. You can classify the CT report for the liver features or symptoms.
User	Now you are going to perform a multi-label classification for a text report of liver computed tomography (CT). Given the potential categorized features {'cyst': 0, 'HCC': 1, 'post-treatment': 2, 'cirrhosis': 3, 'steatosis': 4, 'metastasis': 5, 'hemangioma': 6}, please read the following liver computed tomography report of a patient. If the report is positive with a feature, please return the corresponding value. Liver computed tomography: **[REPORT]**
Assistant Response (Example)	1. HCC (Hepatocellular carcinoma) 3. Cirrhosis

Table 8: Concept Unique Identifiers (CUIs) used with MetaMap (Aronson, 2001) and NegBio (Peng et al., 2018) for our task. RFA: Radiofrequency Ablation; TACE: Transarterial Chemoembolization.

Category	**CUI Term (CUI)**
Cyst	Liver cyst (C0267834)
HCC	Liver carcinoma (C2239176)
Post-Treatment	RFA (C0850292)
	TACE (C3539919)
	Embolization, Therapeutic (C0013931)
	Lobectomy (C0023928)
Cirrhosis	Liver Cirrhosis (C0023890)
Steatosis	Fatty Liver (C0015695)
Metastasis	Metastasis (C4255448)
Hemangioma	Hemangioma (C0018916)