Leveraging Prompt-Learning for Structured Information Extraction from Crohn's Disease Radiology Reports in a Low-Resource Language

Liam Hazan[1], **Gili Focht**[5], **Naama Gavrielov**[2], **Roi Reichart**[1], **Talar Hagopian**[3], **Mary-Louise C. Greer**[4], **Ruth Cytter Kuint**[3], **Dan Turner**[5], and **Moti Freiman**[2]

[1]Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology, Haifa, Israel
[2]Faculty of Biomedical Engineering, Technion - Israel Institute of Technology, Haifa, Israel
[3]Department of Radiology, Shaare Zedek Medical Center, Jerusalem, Israel
[4]Department of Radiology, Hospital for Sick Children, Toronto, Canada
[5]The Juliet Keidan Institute of Pediatric Gastroenterology,
Shaare Zedek Medical Center, Jerusalem, Israel

Abstract

Automatic conversion of free-text radiology reports into structured data using Natural Language Processing (NLP) techniques is crucial for analyzing diseases on a large scale. While effective for tasks in widely spoken languages like English, generative large language models (LLMs) typically underperform with less common languages and can pose potential risks to patient privacy. Fine-tuning local NLP models is hindered by the skewed nature of real-world medical datasets, where rare findings represent a significant data imbalance. We introduce SMP-BERT, a novel prompt learning method that leverages the structured nature of reports to overcome these challenges. In our studies involving a substantial collection of Crohn's disease radiology reports in Hebrew (over 8,000 patients and 10,000 reports), SMP-BERT greatly surpassed traditional fine-tuning methods in performance, notably in detecting infrequent conditions (AUC: 0.99 vs 0.94, F1: 0.84 vs 0.34). SMP-BERT empowers more accurate AI diagnostics available for low-resource languages.

1 Introduction

Medical imaging, particularly Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), emerges as a key element in the management of complex conditions such as Crohn's Disease (CD) (Minordi et al., 2022) serving as a cornerstone for diagnosis, monitoring, and guiding treatment decisions (Bruining et al., 2018). Large-scale analyses of imaging data in CD hold promise for advancing research on the inflammatory burden in the bowel and developing predictive models of disease progression (Gu et al., 2024). The critical clinical information extracted from these images is typically embedded in free-text radiology reports, presenting a significant challenge for large-scale analysis.

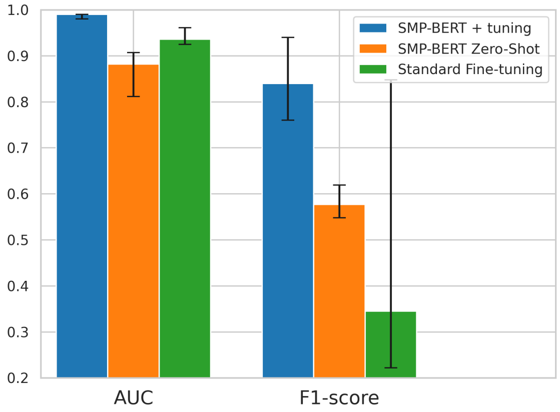
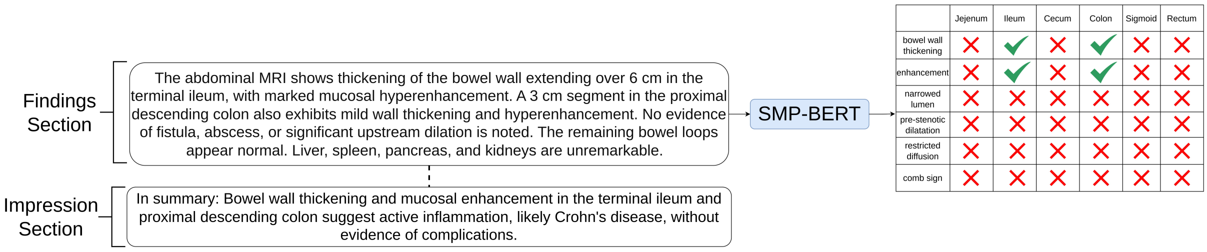


Figure 1: Comparison of the median AUC and F1-score of three models (Standard Fine-tuning, SMP-BERT Zero-Shot, and SMP-BERT + tuning) over all phenotypes with 10+ positives. Error bars represent the Interquartile Range (IQR).

Manually extracting phenotypes and other pertinent information from radiology reports is labor-intensive and requires domain-specific expertise in radiology. Furthermore, CD exhibits high heterogeneity in the disease course, necessitating manual evaluation of a wide range of potential conditions (Torres et al., 2017). This task's time-consuming nature and impracticality for large-scale applications pose significant challenges in achieving efficient and accurate data extraction.

Recent attempts to automate this extraction process have utilized generative Large Language Models (LLMs) such as GPT-4, which leverage free-text instructions instead of requiring annotated data for training (Liu et al., 2023b). While these models hold promise, concerns regarding low-resource languages and data privacy remain a challenge.

Other approaches have involved directly fine-tuning open-source language models on a manually labeled subset of the data (Smit et al., 2020; Yan



The abdominal MRI shows thickening of the bowel wall extending over 6 cm in the terminal ileum, with marked mucosal hyperenhancement. A 3 cm segment in the proximal descending colon also exhibits mild wall thickening and hyperenhancement. No evidence of fistula, abscess, or significant upstream dilation is noted. The remaining bowel loops appear normal. Liver, spleen, pancreas, and kidneys are unremarkable.

In summary: Bowel wall thickening and mucosal enhancement in the terminal ileum and proximal descending colon suggest active inflammation, likely Crohn's disease, without evidence of complications.

	Jejenum	Ileum	Cecum	Colon	Sigmoid	Rectum
bowel wall thickening	✗	✓	✗	✓	✗	✗
enhancement	✗	✓	✗	✓	✗	✗
narrowed lumen	✗	✗	✗	✗	✗	✗
pre-stenotic dilatation	✗	✗	✗	✗	✗	✗
restricted diffusion	✗	✗	✗	✗	✗	✗
comb sign	✗	✗	✗	✗	✗	✗

Figure 2: Example of SMP-BERT Input and Output. A medical radiology report section relevant to a patient's CD diagnosis. The section labeled "Findings" serves as the input for the SMP-BERT model, similar to its pre-training phase.

et al., 2022). However, fine-tuning performance suffers from significant data imbalance, a common challenge in medical datasets and particularly in the case of CD, which features some rare conditions.

To address these limitations, we propose SMP-BERT, a novel prompt learning method built upon the "pre-train, prompt, and predict" framework (Liu et al., 2023a), specifically tailored for the structured nature of radiology reports. SMP-BERT leverages a new pre-training task called Section Matching Prediction (SMP). This task leverages the structured format of radiology reports, where key findings reside in some "Impression" section. By pre-training on this task, SMP-BERT can infer in a zero-shot setting and also further fine-tune using a relatively small amount of annotated data. This approach not only mitigates the challenge of data imbalance but also eliminates the need for massive training corpora during pre-training. This advantage makes SMP-BERT readily applicable to low-resource languages, paving the way for a more inclusive and efficient method of extracting information from radiology reports.

2 Related Work

2.1 Radiology Reports Information Extraction

Various natural language processing approaches have been used in the past to extract information and identify findings on radiology reports, from rule-based methods to deep learning–based language models (Smit et al., 2020; Mozayan et al., 2021; Tejani et al., 2022; Fink et al., 2022). While deep learning models like ClinicalBERT (Huang et al., 2019), and RadBERT (Yan et al., 2022) exploited the use of pre-training on clinical notes and radiology reports, they still require human annotation and a somewhat balanced dataset for fine-tuning.

Generative LLMs, such as GPT-4 and Cluade, may have clear advantages: They don't require extra training and can be easily instructed in natural language to do the task with high performance (Liu et al., 2023b). Unfortunately, radiology reports are usually confidential and can't be sent as a query through the Internet. Although open-source LLMs might be the solution (Mukherjee et al., 2023) they are still focused on English and struggle when it comes to low-resource languages. Moreover, even GPT4 gets comparable results to those of fine-tuned BERT in German (Adams et al., 2023) and an open-source model Vicuna-13B also gets comparable results to BERT-based model (Mukherjee et al., 2023).

2.2 Prompt Learning

Prompt learning (Liu et al., 2023a) is a recent advancement in Natural Language Processing (NLP) that offers a powerful alternative to traditional supervised learning methods which rely on extensive datasets for training a model $P(y|x;\theta)$. Utilizing pre-trained language models (LMs), this approach employs specific input prompts to extend the models' capabilities to tasks beyond their original training. It capitalizes on the input text's probability $P(x;\theta)$, enabling effective use of the comprehensive knowledge amassed by LMs during pre-training. Prompt learning's benefits include its efficient use of data, versatility across different tasks, and reduced need for additional extensive training.

Most prompt learning techniques are based on token-level pre-training tasks such as Left-to-Right Language Modeling (Radford et al., 2019; Brown et al., 2020) or Masked Language Modeling (Schick and Schütze, 2021a,b). However, a handful of approaches operate at the sentence level, such as (Wang et al., 2021), which reformulates the classification task into an entailment task between two sentences.

NSP-BERT (Sun et al., 2022) is another technique that employs sentence-level pre-training through the Next Sentence Prediction (NSP) task. It uses a structured input format beginning with a `[CLS]` token, followed by two sentences, A and B, separated by a `[SEP]` token. The training model balances instances where B genuinely follows A (IsNext) with cases where B is a random sentence (NotNext). The NSP component predicts the likelihood of B following A, relying on a specific matrix W_{nsp} and the `[CLS]` token's hidden vector. For tasks like sentiment analysis, one might use a sentence such as "The ambiance of the restaurant was cozy and inviting," and assess if the sentiment is positive by juxtaposing it with prompts like "The sentiment of this sentence is positive." and "The sentiment of this sentence is negative.", comparing their "IsNext" probabilities. This approach allows labels to correspond with phrases of varying lengths, crucial for extracting information from radiology reports, which often contain findings described in multiple words.

NSP-BERT is optimized for classifying individual sentences, as demonstrated in the pre-training task 3. However, radiology reports consist of multiple sentences, posing a challenge for its application. Furthermore, NSP-BERT capitalizes on the logical progression found in narrative texts, where the sequence of ideas or events aids in making predictions. Contrarily, radiology reports primarily present factual details without a narrative flow, diminishing the method's effectiveness in such contexts.

3 SMP-BERT Framework

3.1 Section Matching Prediction

To overcome these challenges, we propose the Section Matching Prediction (SMP) task, designed specifically for analyzing radiology reports. These reports typically contain structured sections, notably "Findings" and "Impression". The "Findings" segment provides detailed observations from radiological examinations, while the "Impression" segment offers crucial observations and their summarized interpretations. SMP, inspired by the Next Sentence Prediction approach, considers "Findings" as the first segment and "Impression" as the follow-up. During training, "Impression" sections are accurately matched with their "Findings" counterparts half of the time (`Match`), and mismatched the rest (`NotMatch`).

Let $\mathcal{M}$ denote the model trained on our radiology reports. The model is trained on the SMP task where x^F and x^I represent the findings and impression sections, respectively. The model's input takes the following form:

$x_{input} = \texttt{[CLS]}x_i^F\texttt{[SEP]}x_i^I\texttt{[EOS]}$

Let $q_{\mathcal{M}}(n_k|x_i^F, x_i^I)$ denotes the output probability from the model's SMP head based on the input, where $n \in \{\texttt{Match, NotMatch}\}$. The scores s are computed by: $s = \mathrm{W}_{smp}(\mathrm{Tanh}(\mathrm{Wh}_{\texttt{[CLS]}} + \mathrm{b}))$ where $\mathrm{h}_{\texttt{[CLS]}}$ represents the hidden vector of the special token `[CLS]` and W_{smp} is the SMP head matrix. The output probability is calculated using the softmax function:

$$q_{\mathcal{M}}(n_k|x_i^F, x_i^I) = \frac{\exp s(n_k|x_i^F, x_i^I)}{\sum_n \exp s(n|x_i^F, x_i^I)}$$

This training process, optimized by a cross-entropy loss function, allows the model to discern and assess the logical link between these report sections effectively. During inference, we can leverage this learned ability to construct prompts that specifically target the presence or absence of findings in our reports.

3.2 Inference with SMP-BERT

In the inference stage, SMP-BERT leverages its pre-trained understanding of the connection between "Findings" and "Impression" sections. We substitute the "Impression" section with a prompt corresponding to the presence/absence of a clinical finding. By analyzing both the "Findings" section and the prompt, SMP-BERT assigns a higher probability to "`Match`"" when the prompt aligns with the content of the "Findings" section. The input for inference is formulated as: $x_{input} = \texttt{[CLS]}x_i^F\texttt{[SEP]}p^j\texttt{[EOS]}$. Here, p^j represents the prompt corresponding to the j'th label (presence/absence of a finding).

The template $\mathcal{T}$ combines the report's findings section (x_i^F) with generalized prompt: $\mathcal{T}(x) = \texttt{[CLS]}\ x^F\ \texttt{[SEP]}$ There {is/isn't} {finding} in the {organ} `[EOS]`. This approach maps labels to prompts of varying lengths. A verbalizer function $f : \mathcal{Y} \rightarrow \mathcal{P}$ associates each label $y^j \in \mathcal{Y}$ with its corresponding prompt $p^j \in \mathcal{P}$. For example, let $p^j =$ "There is narrowed lumen in the Ileum" and $p^k =$ "There is **not** narrowed lumen in the Ileum" then, the prediction for report x_i regarding narrowed lumen in the Ileum would be $\texttt{argmax}\ (q_{\mathcal{M}}(\texttt{Match}|x_i^F, p^k), q_{\mathcal{M}}(\texttt{Match}|x_i^F, p^j))$.

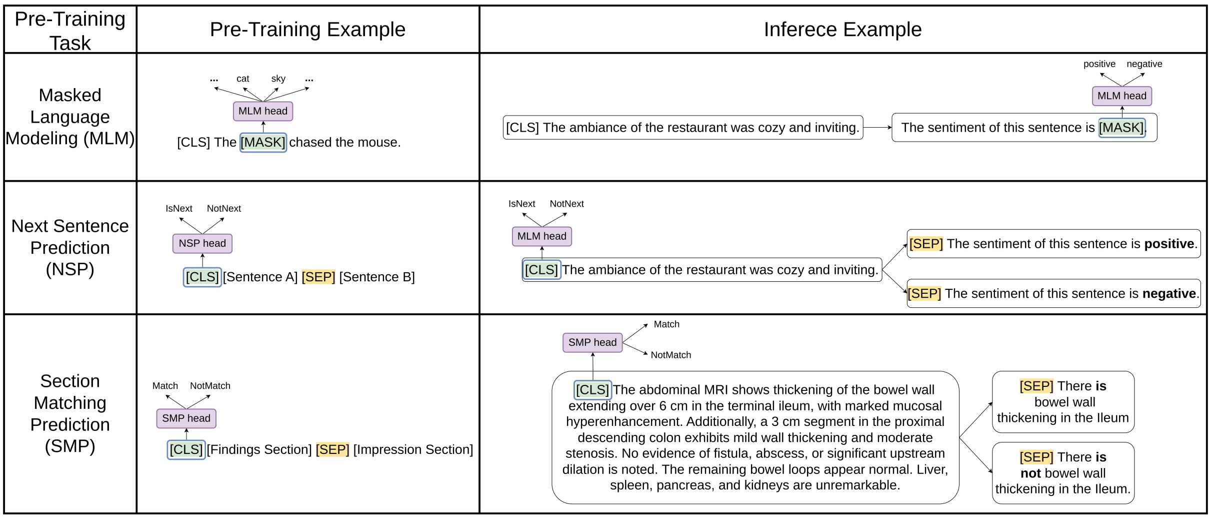

Pre-Training Task
Pre-Training Example
Inferece Example
Masked Language Modeling (MLM)
... cat sky ...
MLM head
[CLS] The [MASK] chased the mouse.
positive negative
MLM head
[CLS] The ambiance of the restaurant was cozy and inviting.
The sentiment of this sentence is [MASK].
Next Sentence Prediction (NSP)
IsNext NotNext
NSP head
[CLS] [Sentence A] [SEP] [Sentence B]
IsNext NotNext
MLM head
[CLS] The ambiance of the restaurant was cozy and inviting.
[SEP] The sentiment of this sentence is positive.
[SEP] The sentiment of this sentence is negative.
Section Matching Prediction (SMP)
Match NotMatch
SMP head
[CLS] [Findings Section] [SEP] [Impression Section]
SMP head
Match
NotMatch
[CLS] The abdominal MRI shows thickening of the bowel wall extending over 6 cm in the terminal ileum, with marked mucosal hyperenhancement. Additionally, a 3 cm segment in the proximal descending colon exhibits mild wall thickening and moderate stenosis. No evidence of fistula, abscess, or significant upstream dilation is noted. The remaining bowel loops appear normal. Liver, spleen, pancreas, and kidneys are unremarkable.
[SEP] There is bowel wall thickening in the Ileum
[SEP] There is not bowel wall thickening in the Ileum.

Figure 3: SMP-BERT Methodology - This figure illustrates three pre-training tasks and how they can be used for text classification through prompt learning. Using MLM (token-level) for inference requires "cloze question" prompts and a verbalizer function to convert labels into single-token answers (e.g., "positive"/"negative"). Using NSP (sentence-level) is more simple. While it allows prompts of varying lengths, it's still limited to single-sentence classification. Our novel SMP solves it by pre-training on matching whole sections (multiple sentence level). Then, replace the "Impression" section with a prompt about the presence/absence of a finding.

3.3 SMP-tuning

The SMP-tuning process is visualized in Figure 4 and conducted similarly to the approach of NSP-tuning from NSP-BERT (Sun et al., 2022).

Generally, this process is a continuation of the SMP pre-training just given annotated reports we use the prompts instead of actual "Impression" sections. Given a sample i with its reference label y_i^+, we define a positive instance as $(\mathcal{T}(x_i, y_i^+), \texttt{Match})$ and for each label y_i^- that does not match the reference label, we define negative instances as $\{(T(x_i, y_i^-), \texttt{NotMatch})\}_{y_i^- \in Y \setminus \{y_i^+\}}$, where Y is the set of all possible labels. This constructed data sums up to (n_samples*n_phenotypes*n_labels) instances and then used to fine-tune the model, leveraging the initialized weights from the SMP pre-training phase.

4 Experiments

4.1 Data

This study's dataset consists of radiology reports from three medical institutions, spanning 2010 to 2023. This dataset contains 9,683 free-text reports (one for each visit) for 8093 distinct patients. Since this dataset is confidential, no study has used it to assess the performance of any model. Ethics approval was obtained from the Shaare Zedek Medical Center Institutional Review Board (Helsinki) committee.

For this study, a subset of 700 reports were manually annotated for the presence or absence of certain phenotypes in various organs according to the Consensus Recommendations of the American Gastroenterological Association and the Society for Abdominal Radiology (Bruining et al., 2018). The annotations focused on the following organs: organs jejunum, ileum, cecum, colon, sigmoid, and rectum. Specific findings annotated included bowel wall thickening, hyper-enhancement, pre-stenotic dilatation, narrowed lumen, restricted diffusion, and comb sign. Since our radiology reports are in the form of free text, we segmented them into "Findings" and "Impression" sections using keywords like "In summary:".

4.2 Experimental Setup

We divided the dataset into three distinct sets using a multi-label stratification (Sechidis et al., 2011): training (300 reports), validation (100 reports), and test (300 reports) as illustrated in Figure 5. This stratification was crucial to maintain representative distributions of labels across the sets, considering the significant class imbalance present in the majority of labels.

Our goal was to compare the performance of our method against standard fine-tuning and assess the advantages of adding the SMP-tuning step on top of the zero-shot approach.

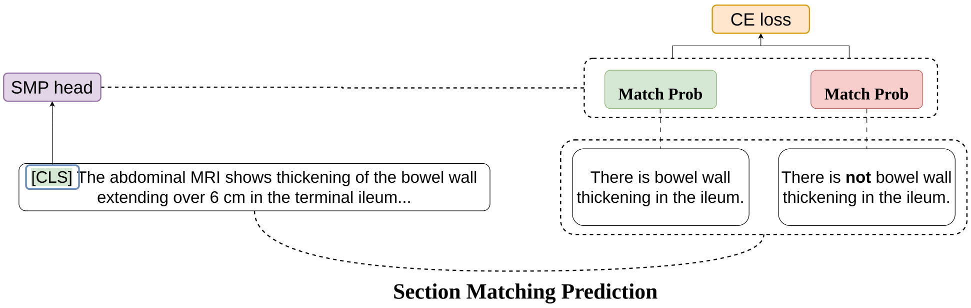

CE loss
SMP head
Match Prob
Match Prob
[CLS] The abdominal MRI shows thickening of the bowel wall extending over 6 cm in the terminal ileum...
There is bowel wall thickening in the ileum.
There is **not** bowel wall thickening in the ileum.
Section Matching Prediction

Figure 4: SMP-tuning - Fine-tuning SMP-BERT by generating a negative and a positive instance for every annotated sample and every label. The true label is "There is finding ..." so the negative instance is paired with "There is not finding ..."

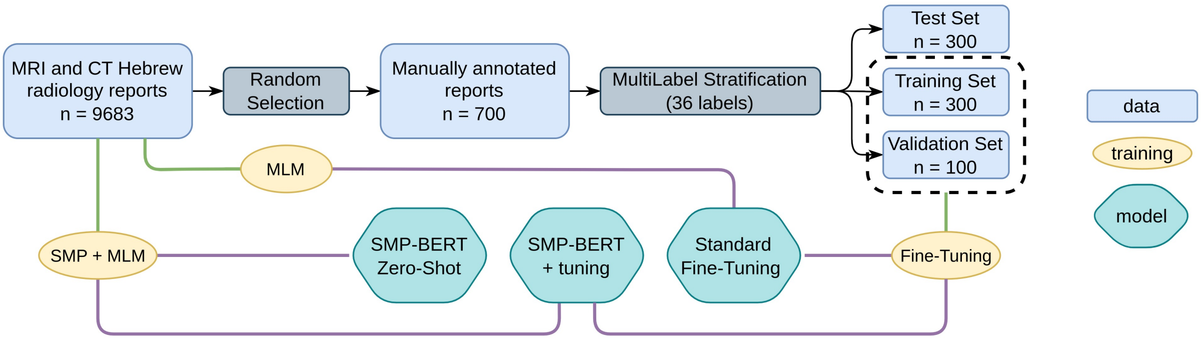


Figure 5: Flowchart of study design - The flowchart outlines the sequence of processing steps from data acquisition to model evaluation. It visualizes the progression from the initial collection of MRI and CT Hebrew radiology reports, through the stages of manual annotation and multi-label stratification, culminating in the pre-training/training of the different models.

The foundation of our models is the Hebrew RoBERTa (HeRo) model (Shalumov and Haskey, 2023), initially pre-trained on the HeDC4 corpus, a comprehensive Hebrew language corpus. We further pre-trained the model on all our radiology reports using the Masked Language Modeling (MLM) task, since there are no other open medical large corpora for Hebrew.

We conducted experiments using three models:

- **Standard Fine-tuning**: This model was fine-tuned directly for multi-label classification for all phenotypes.
- **SMP-BERT Zero-Shot**: This model was further pre-trained on all radiology reports using the SMP task. Inference was executed using the SMP-BERT methodology mentioned in the Inference section.
- **SMP-BERT + tuning**: Like the zero-shot model, this model underwent pre-training with the SMP task on all radiology reports. Additionally, it was trained further using SMP-tuning to optimize its performance.

In addition, we assessed the impact of training set size: The models were trained on datasets of varying sizes (50 to 300 reports) to analyze how the amount of training data affects their performance and ability to generalize to unseen data. We further conducted an ablation study to asses the contributions of MLM and SMP pre-training tasks to the model's performance.

Our initial goal was to compare our method with open-source generative LLMs like Llama 2. However, currently available open-source LLMs are not optimized for low-resource languages such as Hebrew, which made the comparison infeasible.

Due to the inherent class imbalance in the dataset, where most labels have a low number of positive samples, we primarily evaluated the models using the F1-score alongside the AUC metric. The F1-score considers both precision and recall, making it well-suited for imbalanced datasets. Additionally, we reported the Interquartile Range (IQR) along with the scores to provide insight into the variability and distribution of model performance across different labels.

All experiments were conducted using a single NVIDIA RTX A6000 GPU, with each experiment taking approximately 1-3 hours.

Hyper-parameters

For SMP-BERT + tuning, we train 6 epochs on the constructed dataset ($300 * 36 * 2 = 21600$). For standard Fine Tuning, we trained 120 epochs on the original data (300). For both we set learning rate as 2e-5 with linear decay and the batch size is 24.

5 Results

To account for the inherent class imbalance in our dataset, we focused our analysis on phenotypes with at least 10 positive samples, ensuring the reliability of our findings.

Our evaluation across three distinct model configurations highlighted the superior performance of the SMP-BERT + tuning approach in extracting phenotypic information from CD radiology reports. The SMP-BERT + tuning model achieved the highest median AUC of 0.99 (IQR 0.98-0.99), outperforming the Standard Fine-tuning model's median AUC of 0.94 (IQR 0.92-0.96) and the SMP-BERT Zero-Shot model's median AUC of 0.88 (IQR 0.81-0.91). For F1-score evaluations, the SMP-BERT + tuning model again leads with a median score of 0.84 (IQR 0.76-0.94), which is substantially higher than the scores of the Standard Fine-tuning model (0.34, IQR 0.22-0.85) and the SMP-BERT Zero-Shot model (0.58, IQR 0.55-0.62). A comprehensive breakdown of these results, including F1 and AUC scores for individual phenotypes, is detailed in the accompanying Table 1.

Further analysis presented in Figure 7 of model performance relative to the count of positive instances exhibited the strength of SMP-BERT + tuning, particularly for labels with sparse positives in the training set. For example, with only 19 positive cases for "Rectum Bowel Wall Thickening," SMP-BERT + tuning achieved a significantly higher F1-score (0.74) compared to the standard model (0.1). This demonstrates its superior ability to generalize well from limited data.

However, both models performed well when dealing with abundant positive instances. For example, with 137 positives for "Ileum Bowel Wall Thickening" (almost half the dataset), both models achieved good results, with SMP-BERT + tuning maintaining a decent gap (F1-score 0.97 vs. 0.915 for the standard model).

The graph shown in Figure 7 indicates that the performance gap between the models decreases with an increase in the number of positive instances. This suggests that while SMP-BERT + tuning shines with limited data, it still performs better when more data is available.

We also analyzed how the size of the training set impacts model performance. As shown in Figure 6, the SMP-BERT + tuning model exhibits superior adaptability. Notably, it achieves good performance even with limited training data (50-100 samples). The Standard Fine-tuning model exhibits a trend of broadening IQRs and decrease of median score. This could suggests an improving performance for common phenotypes (like Ileum Bowel Wall Thickening) but potentially decreasing performance for rarer ones due to increased data imbalance.

Ablation Study

As evidenced by Table 2, both pre-training tasks, MLM and SMP, significantly contribute to optimizing the performance of SMP-BERT. Moreover, it appears that standard fine-tuning benefits from the inclusion of the SMP task.

6 Discussion

This study examined the efficacy of SMP-BERT, a novel prompt-learning approach, in extracting detailed information from Hebrew radiology reports of CD patients. Our results reveal that SMP-BERT, especially the fine-tuned version (SMP-BERT + tuning), significantly outperforms the standard fine-tuning approach, , achieving an improvement of 49% in median F1 score and 5% in median AUC.

Our study highlights the significant improvement of SMP-BERT + tuning, achieving superior F1-scores and AUCs compared to standard fine-tuning across all analyzed phenotypes. Notably, the model performs well even with a low amount of annotated data. This improvement is particularly notable for rarer phenotypes, demonstrating the model's ability to handle imbalanced datasets, a common challenge in the medical domain. This robustness is crucial for advancing research in CD and other conditions with diverse clinical presentations.

Furthermore, this study contributes to the growing exploration of prompt learning for NLP tasks in healthcare. Unlike traditional fine-tuning approaches, which require substantial labeled data, SMP-BERT leverages pre-training on the “Section Matching Prediction” task and further SMP-tuning

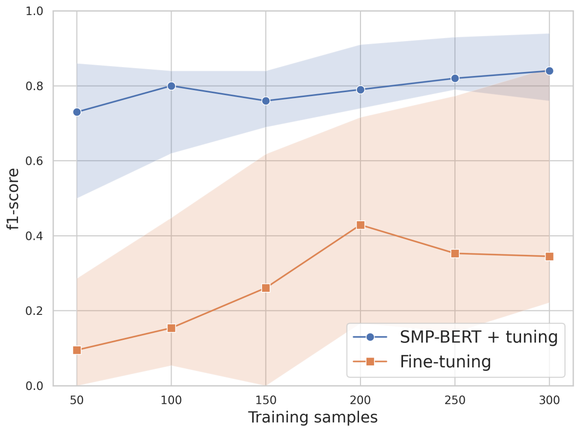


Figure 6: Median F1 scores and IQRs for SMP-BERT + tuning and Standard fine-tuning trained on different training set sizes.

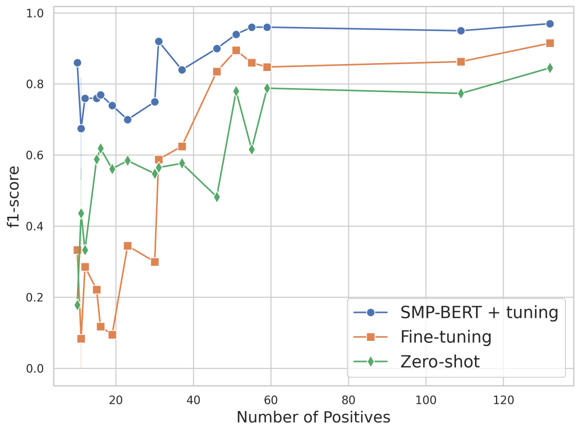


Figure 7: This line chart plots the F1 scores against the number of positive instances of all phenotypes in the dataset (300 total).

to achieve exceptional performance even with limited data. This opens exciting possibilities for applying prompt learning in scenarios with limited annotated data, imbalanced data, or low-resource languages, pushing the boundaries of NLP applications in healthcare.

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).

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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

David H. Bruining, Ellen M. Zimmermann, Edward V. Loftus, William J. Sandborn, Cary G. Sauer, and Scott A. Strong. 2018. Consensus recommendations for evaluation, interpretation, and utilization of computed tomography and magnetic resonance enterography in patients with small bowel crohn's disease. *Radiology*, 286(3):776–799.

Matthias A. Fink, Klaus Kades, Arved Bischoff, Martin Moll, Merle Schnell, Maike Küchler, Gregor Köhler, Jan Sellner, Claus Peter Heussel, Hans-Ulrich Kauczor, Heinz-Peter Schlemmer, Klaus Maier-Hein, Tim F. Weber, and Jens Kleesiek. 2022. Deep learning–based assessment of oncologic outcomes from natural language processing of structured radiology reports. *Radiology: Artificial Intelligence*, 4(5).

Phillip Gu, Oreen Mendonca, Dan Carter, Shishir Dube, Paul Wang, Xiuzhen Huang, Debiao Li, Jason H Moore, and Dermot P B McGovern. 2024. Ailuminating artificial intelligence in inflammatory bowel diseases: A narrative review on the role of ai in endoscopy, histology, and imaging for ibd. *Inflammatory Bowel Diseases*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya Nori, Matthew Lungren, Ozan Oktay, and Javier Alvarez-Valle. 2023b. Exploring the boundaries of gpt-4 in radiology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Laura Maria Minordi, Antonio Bevere, Alfredo Papa, Luigi Larosa, and Riccardo Manfredi. 2022. Ct and mri evaluations in crohn's complications: A guide for the radiologist. *Academic Radiology*, 29(8):1206–1227.

Organ-Finding	**SMP-BERT + tuning**	**SMP-BERT Zero-Shot**	**Standard Fine-Tuning**	**prevalence**
ileum-bowel wall thickening	**0.97/1.0**	0.85/0.91	0.92/0.98	44%
ileum-enhancement	**0.95/0.99**	0.77/0.86	0.86/0.95	36%
ileum-narrowed lumen	**0.96/1.0**	0.79/0.92	0.85/0.97	19%
ileum-dilatation	**0.96/1.0**	0.62/0.87	0.86/0.96	18%
ileum-comb sign	**0.9/0.99**	0.48/0.81	0.84/0.97	15%
ileum-restricted diffusion	**0.94/0.99**	0.78/0.91	0.9/**0.99**	16%
colon-bowel wall thickening	**0.84/0.98**	0.58/0.88	0.62/0.93	12%
colon-enhancement	**0.92/0.99**	0.57/0.88	0.59/0.95	9%
colon-comb sign	**0.86/1.0**	0.18/0.74	0.33/0.94	3%
colon-restricted diffusion	**0.76/0.98**	0.33/0.94	0.29/0.91	3%
rectum-bowel wall thickening	**0.74/0.96**	0.56/0.89	0.1/0.96	6%
rectum-enhancement	**0.76/0.98**	0.59/0.78	0.22/0.89	5%
sigmoid-bowel wall thickening	**0.75/0.97**	0.55/0.77	0.3/0.9	10%
sigmoid-enhancement	**0.7/0.98**	0.58/0.89	0.34/0.88	7%
sigmoid-comb sign	**0.53/0.98**	0.31/0.78	0.17/0.93	3%
cecum-bowel wall thickening	**0.77/0.98**	0.62/0.89	0.12/0.93	5%
cecum-enhancement	**0.82/0.99**	0.56/0.93	0.0/0.92	3%

Table 1: Performance comparison. Values are F1/AUC scores for each model across different phenotypes. The Prevalence column indicates the percentage of test samples in which the phenotype is present.

Method	MLM	SMP	F1-Score	AUC
SMP-BERT + tuning	✓	✓	**0.84 [0.76,0.94]**	**0.99 [0.98,0.99]**
	✓	×	0.75 [0.59,0.87]	0.97 [0.96,0.98]
	×	✓	0.73 [0.67,0.89]	0.97 [0.95,0.98]
	×	×	0.42 [0.26,0.57]	0.94 [0.92,0.96]
Standard Fine-tuning	✓	✓	**0.55 [0.35,0.86]**	**0.96 [0.95,0.98]**
	✓	×	0.34 [0.22,0.85]	0.94 [0.92,0.96]
	×	✓	0.15 [0.0,0.72]	0.85 [0.82,0.91]
	×	×	0.12 [0.0,0.61]	0.83 [0.78,0.88]

Table 2: Ablation Study on Pre-training Tasks.

Ali Mozayan, Alexander R. Fabbri, Michelle Maneevese, Irena Tocino, and Sophie Chheang. 2021. Practical guide to natural language processing for radiology. *RadioGraphics*, 41(5):1446–1453.

Pritam Mukherjee, Benjamin Hou, Ricardo B. Lanfredi, and Ronald M. Summers. 2023. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*, 309(1).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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*. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. 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*. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. *On the Stratification of Multi-label Data*, page 145–158. Springer Berlin Heidelberg.

Vitaly Shalumov and Harel Haskey. 2023. Hero: Roberta and longformer hebrew language models. *arXiv preprint arXiv:2304.11077*.

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 Meth-*

ods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

Yi Sun, Yu Zheng, Chao Hao, and Hangping Qiu. 2022. NSP-BERT: A prompt-based few-shot learner through an original pre-training task —— next sentence prediction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3233–3250, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ali S. Tejani, Yee S. Ng, Yin Xi, Julia R. Fielding, Travis G. Browning, and Jesse C. Rayan. 2022. Performance of multiple pretrained bert models to automate and accelerate data annotation for large datasets. *Radiology: Artificial Intelligence*, 4(4).

Joana Torres, Saurabh Mehandru, Jean-Frédéric Colombel, and Laurent Peyrin-Biroulet. 2017. Crohn's disease. *The Lancet*, 389(10080):1741–1755.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4).