Prompt-Eng: Healthcare Prompt Engineering

Revolutionizing Healthcare Applications with Precision Prompts

Awais Ahmed
202014080105@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Mengshu Hou
mshou@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Rui Xi*
ruix2022@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Xiaoyang Zeng
202011081605@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Syed Attique Shah*
syedattique.shah@bcu.ac.uk
Birmingham City University
Birmingham, United Kingdom

ABSTRACT

Prompt Engineering has emerged as a pivotal technique in Natural Language Processing, providing a flexible approach for leveraging pre-trained language models. Particularly, a prompt is used to instruct the model to adopt the nature of given prompts, which became a well-adoptable approach in wide areas of domains. Yet, existing prompt-guided frameworks are experiencing various challenges, such as crafting prompts for specific tasks to achieve clarity and conciseness and avoid ambiguity, which requires time and computational resources. Further existing methods heavily rely on the extensive labelled datasets, yet many domain-specific challenges exist, particularly in healthcare. This study presents Prompt-Eng, a novel framework emphasizing its wide-ranging applications in healthcare, where we design precise prompts with positive and negative aspects; we hypothesize that designing prompts in pairs helps models to generalize effectively. We delve into the significance of quick design and optimization, highlighting its influence in shaping model responses. In addition, we explore the increasing demand for prompts that are aware of the context in multimodal data analysis and the incorporation of prompt engineering in new machine-learning approaches. The essence of our approach is to create tailored prompts, which serve as instructive guidelines for the models during the prediction procedure. The proposed methodology emphasizes utilizing context-aware prompt pairs to facilitate interpreting and extracting healthcare information from a health corpus by models. The study uses the medical MIMIC-III [1] corpus to predict medicine prescriptions. The paper also explores visual and textual prompts for X-ray image analysis for pneumonia prediction on the MIMIC-CXR [2] dataset. This approach stands out from existing methods by addressing challenges such as clarity, conciseness, and context awareness, thereby enabling improved interpretation and extraction of healthcare information from diverse data sources.

CCS CONCEPTS

• **Information systems** → **Information systems applications**.

KEYWORDS

Prompt-Engineering, Healthcare Prompts, Healthcare Analysis, Natural Language Processing, Precision Healthcare.

ACM Reference Format:
Awais Ahmed, Mengshu Hou, Rui Xi, Xiaoyang Zeng, and Syed Attique Shah. 2024. Prompt-Eng: Healthcare Prompt Engineering: Revolutionizing Healthcare Applications with Precision Prompts. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3589335.3651904

1 INTRODUCTION

In recent years, integrating artificial intelligence (AI) and natural language processing (NLP) technologies into healthcare systems has significantly advanced medical diagnostics, treatment recommendation systems, and patient care [1]. One of the key challenges in this domain is the accurate interpretation and utilization of vast amounts of medical data to make informed decisions [2]. To address this challenge, prompt engineering has emerged as a promising approach to enhance the performance and interpretability of AI models in healthcare applications [3–5]. Prompt engineering involves designing precise and contextually relevant prompts that guide AI models to generate accurate predictions or analyses [6].

A "prompt" [3] can considered an additional contextual natural language sentence or phrase fed to an AI model, which helps to activate models and provoke certain responses to the tasks [7]. Prompts could play a crucial role in healthcare and help extract contextual guidance and explicit instructions to AI models, aiding in the accurate extraction of medical information and improving prediction performance. One notable use is in medical image analysis, where

*Corresponding author
[1]https://physionet.org/content/mimiciii/1.4/

[2]https://physionet.org/content/mimic-cxr/2.0.0/

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore.

ACM ISBN 979-8-4007-0172-6/24/05…$15.00
https://doi.org/10.1145/3589335.3651904

prompts can guide AI models to focus on specific regions of interest or pathologies within images. For example, by providing a prompt highlighting the presence of pulmonary nodules in chest X-rays, AI models can be trained to detect and classify these abnormalities more accurately. Prompts can also assist clinical decision support systems by contextualizing patient data and providing relevant information to healthcare providers [8]. By incorporating prompts that capture pertinent clinical details from electronic health records, AI systems can generate more personalized treatment recommendations tailored to individual patient needs.

In healthcare, the effectiveness of precise prediction plays a critical role. AI-driven healthcare systems demand time, computing resources, extensive labeled data, and an aim toward scalability. When designing Prompt-Eng framework, we conducted a comprehensive literature review to identify relevant studies that cover broader context healthcare applications using the prompts technique. Still, none of the studies have presented a comprehensive analysis. Further, we notice the following challenges that warrant a state-of-the-art healthcare system to achieve the core objectives that necessitate the utilization of prompt engineering.

- Developing time and resource-efficient solutions.
- Avoiding reliance on training-intensive techniques
- Extensive labeled data is needed
- Addressing domain-specific challenges arising from the dynamic nature of healthcare.

To address the above challenges and fill the research gap, this paper introduces Prompt-Eng, a novel framework that leverages prompt engineering techniques to revolutionize healthcare applications. By integrating context-aware prompts with state-of-the-art AI models, Prompt-Eng aims to improve diagnostic accuracy, treatment recommendations, and overall patient outcomes. Further, our prompt engineering framework is tailored to achieve precise healthcare results within a broader context.

In summary, the main contributions of this work are as follows:

- ***Healthcare-Aware Prompt Design:*** The study suggests crafting precise paired prompts for a wide range of healthcare tasks.
- ***Prompt-Eng:*** The study presents a Prompt-Eng framework that utilizes pre-trained models, incorporating the core of transformers.
- ***Wide Applications In Healthcare:*** The study first develops a theoretical background and highlights the extensive applicability of prompt engineering in the healthcare domain, addressing critical challenges and offering tailored solutions.
- ***Experimental Validation:*** We demonstrates the experimental results of the initial hypothesis. This suggests the effectiveness of the proposed approach, providing empirical evidence of the practicality of prompt engineering in healthcare applications.

The sections are organized as follows: Section. 2 brief literature review, and after that, Section. 3 discusses the methodology. Followed by Section. 4 delves into the detailed core health prompt engineering. Finally, the discussion, limitations, and conclusion are presented in Section. 6 to Section. 8, respectively.

2 LITERATURE REVIEW

This section highlights the essential studies in which prompt engineering has assisted healthcare applications or presented a detailed investigation of such methodologies that would benefit the audience.

Prompt engineering can be pivotal in healthcare, similar to applications such as few-shot learning for chronic diseases [9], a real-time EHR question answering framework [10], where authors emphasize utilization of prompt engineering using GPT-4. The study [11] presents a study where they enhance the methodology of summarizing automated medical reporting with the utilization of transformer-based prompt engineering. Another study by [12] examined the feasibility of prompt learning for clinically meaningful tasks utilizing frozen models. The work [13] presents an innovative way for clinical decision-making using OpenAI's ChatGPT to leverage LLMs. Their approach extracts features and important insights that improve decision-making by considering ML models as medical experts. This domain transfer improves diagnostic tools. They also examine LLM-based zero-shot and few-shot prompt learning dynamics. They further compared OpenAI's ChatGPT with classic supervised ML models in diverse data settings to assess the efficacy of quick engineering techniques under different data scenarios. The study [6] introduces a catalog of prompt patterns to improve interaction with ChatGPT, providing a systematic approach to optimizing prompt engineering for various tasks. Given its comprehensive analysis and potential practicability, this study serves as a baseline or fundamental to our work, as our intention is similar, where we also want to emphasize potential prompts in the wider healthcare application.

The study [14] presents a review of large vision models and also emphasizes visual prompt engineering. Further, study [15] presents a review study on prompt engineering techniques for ChatGPT, where authors outlined methodologies, recommendations, and optimal approaches. Another review study [16] where authors targeted the potential of prompt engineering. Furthermore, the study [17] presents a detailed tutorial study where they mentioned the potential and importance of prompt engineering as a promising avenue and needed skill for medical practitioners.

The Importance of Pre-Trained Language Models: Language models are primarily designed to predict the next token in a sequence based on a given set of continuous tokens [18]. These models, also known as probabilistic models, assign probabilities to various sequences to generate predictions for the next token. Token prediction is considered a classical application in NLP, and based on this foundational application, numerous language models have been developed in the past. Prompt engineering leverages these models, enhancing their effectiveness across various tasks. This paper introduces Prompt-Eng, a framework for precision prompt engineering in healthcare. It explores prompt optimization strategies and the integration of prompts into models for medical tasks.

3 METHODOLOGY

The core of our methodology consists of two key factors: which are i) Prompt Engineering for healthcare in broader perspectives of application, which aims to develop a theoretical background (this involves the development of a comprehensive theoretical background to understand the principles and applications of prompt engineering in healthcare settings), while the second ii) Using a set of prompts for instructing the model, aiming to enhance model performance and robustness. The model architecture diagram, as depicted in Figure 1, is redesigned from our original publication [4], where we only focused on medicine predictions using the clinical

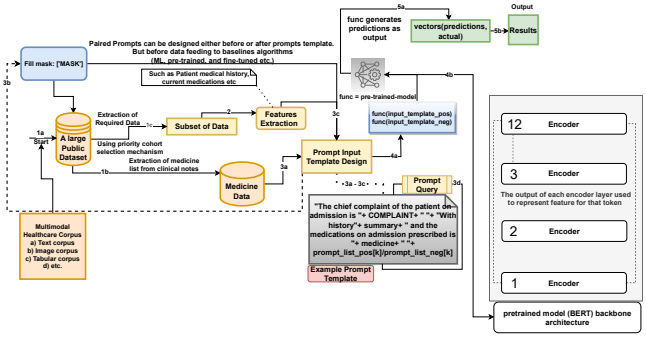


Figure 1: A Proposed Medicine-Prompt Framework: A Systematic Approach to Medicine Prediction from Clinical Notes using Advanced Prompt Engineering and Pre-trained Language Models. - An example model architecture framework that could be adapted for any of the listed healthcare tasks. The architecture is delineated with easy steps, facilitating comprehension and execution.

corpus. In this study, we extend our work by emphasizing prompt engineering applicability to broader perspectives of healthcare. The model architecture is designed to act as a backbone for multimodal healthcare applications where different types of corpus exist. Currently, our corpus is MIMIC-III (tabular data and raw clinical notes), multimodal data perspectives are fed to the model, and relevant prompts are passed to instruct the model. We have performed several preprocessing efforts to generate a meaningful subset of clinical text data that contains various key factors for clinical decision process applications, such as the list of medicines at various stages, information on patients' symptoms, chief complaints of patients, and a summary of patients and patients demographic information. Initially, we put our efforts into proposing an Algorithm. 1 that helped us to generate task-specific prompts. Then, we outlined high-level pseudo steps for developing methodologies for broader healthcare applications, as shown in Algorithm. 2. Then, a fine-tuned approach is employed for conducting detailed experiments for the medical diagnostic categories, particularly medicine prediction, using prompt engineering, as shown in Algorithm. 3.

In summary, the presented methodology is employed for predicting medication from clinical notes, integrating prompt engineering with pre-trained language models. This architecture aims to efficiently parse and interpret clinical text to predict medications, showcasing a novel application of NLP techniques in healthcare. The presented methodology can be employed in broader perspectives of healthcare to leverage the power of prompt engineering and pre-trained models; a few applications are suggested in subsequent Section. 4.

3.1 Preparing Datasets

In this work, we meticulously conducted several processing steps to prepare the dataset for analysis. We accessed the MIMIC-III [19] and MIMIC-CXR [20] datasets for different experiments, adhering to official guidelines and legal prerequisites, including HIPAA

Algorithm 1: A Generalized Template for Generating Prompts for Categories as specified in Section. 4

```
function GeneratePrompts(n,categories,templates) for each catg in categories do
    prompts ← [ ] for i ← 1 to n do
        (template ← RANDOM(templates[category]))
        args ← GENERATESPECIFICS(template)
        prompt ← FORMAT(template, args)
        APPEND prompt to prompts
    end
    PRINT "Category: category, Prompt: prompt"
end
```

Algorithm 2: A High-Level Algorithmic Representation for Generalized Healthcare Tasks using Prompt Engineering.

```
input  : Patient data (e.g., symptoms, medical history, medical images, disease images, etc.)
output : Task-oriented predictions or recommendation, interpretation or classification
Generate Prompts
Collect input data
Formulate prompts
Feed prompts to language models
Evaluate model outputs
Record and Return results
```

Algorithm 3: Prompt-Eng A High-Level Algorithmic Representation of Proposed Approach using Transformer's Model Transfer Learning

```
input  : modeltype, tokentype, m_count, input_data_samples, epochs
output : Predictions Vectors
results_list                ▷ Initializing empty lists, matched_medicines, actual_result, predictions
for i ← 1 to epochs do
    for i ← 1 to input_data_samples do
        Inputs:                                   ▷ Here, several inputs can be considered
        (complaint)
        (medications)
        (summary)
        inputs concatenated
        clear the matched_medicines list
        for each medicine in medicine_count do
            if medicine.lower() is in medications.lower() then
                Append medicine to matched_medicines
            end
        end
        Join the elements of matched_medicines with '|' and assign it to
          matched_medicines_pattern
        for each medicine in medicine_count do
            Initialize lists:
            prompt_list_pos and prompt_list_neg;
            for each prompt in prompt_list_pos and prompt_list_neg do
                Feed the input_text (pos and neg) separately to the Model for
                  training
                call to verbalize_score()
                Comparing scores to determine the prediction
                Calculating loss
                Adjusting training-loss
                loss.backward()                   ▷ loss backward
                optim.step()                      ▷ optimizer
                if score_pos > score_neg then
                    Set prediction to 1 (Positive);
                end
                else
                    Set prediction to 0 (Negative);
                end
                actual_result;
            end
        end
        Append a dictionary to results_list containing the 'Predicted Values' and
          'Actual Values' lists
    end
end
return Output Vectors
```

regulations. Firstly, we accessed and preprocessed MIMIC-III's NOTEEVENTS file, which contains approximately 2.1 million rows and 11 columns of clinical notes in free-text format. We extracted

relevant features from these notes, focusing on key aspects such as patient demographics, services provided, allergies, diseases, major complaints, and medication history. Through a process of keyword filtering, we narrowed down the dataset to include only records containing predefined keywords such as "Chief Complaint," "Medications on Admission," "Discharge Medications," and "History of Present Illness." This meticulous filtering resulted in a subset of 36,000 records suitable for further analysis. Additionally, to enhance the interpretability of the data, summaries of the "History of Present Illness" feature were generated using a pre-trained T5 model. These preprocessing steps ensured that the dataset was appropriately curated and ready for subsequent analysis within the context of our study. For the MIMIC-CXR experiment, we provided x-ray images and preprocessed them on the fly while utilizing model backbones (VGG16 and VGG19) or proposed Prompt-EngCNN.

4 HEALTHCARE PROMPT ENGINEERING

Prompt engineering provides a flexible and adjustable framework for effectively dealing with various issues and possibilities in the healthcare sector. Prompts facilitate extracting significant insights and predictions from clinical data by offering informative instructions to language models. This, in turn, contributes to enhancing patient care and medical decision-making.

Given an input x, traditional supervised learning models predict an output y as follows [3]:

$$P(y|x) : \text{Model predicts output } y \text{ given input } x. \tag{1}$$

In prompt-based learning, the process involves transforming the original input x into a modified textual prompt x' and then filling in the placeholders to obtain the final [MASK/Z] string $\hat{x}$, from which the desired output y is inferred [following equations represents the two steps as per existing literature [21] i) Applying Template ii) Filling Slot]:

$$X \xrightarrow{\text{template}} X' : \text{Input } X \text{ transformed to textual prompt } X' \text{ [Z]}. \tag{2}$$

$$X' \xrightarrow{\text{LM}} \hat{x} : \text{LMP fills the unfilled slots in } X' \text{ to obtain } \hat{x}. \tag{3}$$

$$\hat{X} \rightarrow y : \text{The final output } y \text{ is derived from [Z]}\hat{x}. \tag{4}$$

$$ImgCl : \text{Given an X-Ray image } [X], \text{predict the likelihood of } [Z] \tag{5}$$

$$DiagPred : \begin{array}{r}\text{Given the patient's symptoms (X) of} \\ [X]\text{predict the most likely diagnosis } [Z]\end{array} \tag{6}$$

$$TrtRec = \begin{array}{r}\text{Considering the patient's symptoms and} \\ \text{previous responses to treatment } [X], \\ \text{propose an effective therapy } [Z]\end{array} \tag{7}$$

Subsequent categories delve into broader healthcare application areas, showcasing various areas where prompt engineering could offer substantial benefits.

(a) Medical Diagnosis Prediction: Employing prompts to instruct models in predicting potential diagnoses by analyzing symptoms and clinical narratives documented in patient data. Several experiments are conducted to validate this category, as shown in Table. 2 and Table. 3.
(b) Treatment Recommendation: Using specific prompts, providing treatment plans, or prescribing drugs based on analyzing the patient's medical history, present symptoms, and other relevant clinical criteria.
(c) Patient Triage and Prioritization: Facilitating patient triage by assessing the urgency and severity described in clinical notes, hence assisting in the priority of patient treatment.
(d) Medical Image Interpretation: It involves the analysis of medical imaging, such as X-rays or MRI scans, by utilizing prompts to assist models in recognizing and comprehending crucial visual characteristics pertinent to making a diagnosis.
(e) Clinical Trial Matching: This involves analyzing patient data and identifying appropriate clinical trials by utilizing sophisticated prompts to match the patient's information with the specific requirements of the studies.
(f) Personalized Health Suggestions: Creating customized recommendations for lifestyle or health enhancements by analyzing individual health records and patterns using prompt-guided models.
(g) Automated Medical Coding: Improving medical coding accuracy from clinical texts for billing and administrative purposes by using prompts to extract and categorize pertinent information.
(h) Drug Interaction and Side Effects Analysis: Anticipating probable medication interactions and side effects by examining medical literature and patient data using well-designed prompts.
(i) Mental Health Analysis: The objective is to detect indications of mental health conditions, such as depression or anxiety, by examining the language and behavior of patients as documented in clinical records. This is achieved by employing prompts sensitive to the context in which they are used.
(j) DNA sequences Data Interpretation: It involves analyzing and understanding genome data to provide tailored medical guidance. This is achieved by using prompts to guide models in identifying significant genetic markers and determining their consequences.

The efficiency and usability of prompt engineering would benefit by incorporating it with various techniques and approaches, such as:

- Pre-trained language models (PLMs)
- Transfer learning
- Multi-task learning
- Data augmentation
- Adversarial training
- Contextual embedding

By integrating prompt engineering with the listed approaches, academics and practitioners may create more robust, efficient, and flexible frameworks for leveraging the power of existing models, as outlined in Algorithm. 2. The definitive steps are outlined in an algorithm. Based on this generalized algorithm, we proposed a robust and fine-tuned version for Prompt-Eng utilizing a pre-trained ClinicalBERT model. Fined-tuned (training, validation, and testing) loss results are shown in Figure. 3 for two random medicines for 50 epochs. Additionally, we conducted one more experiment for predicting the random collection of medicine sets with two different prompting techniques, as shown in Figure. 4. These results demonstrate the effectiveness and stability of our approach in optimizing model performance by employing prompts within a clinical corpus.

The table (Table 1) displays a wide range of positive and negative prompts specifically created for various categories of tasks for Prompt-Eng framework. These prompts are crucial in guiding

Table 1: Set of Positive and Negative Prompts for [GPT and Manual] Designed for Prompt-Eng Framework. We have prepared several prompts with the fine-tuned model; we used ⋯ to skip a few prompts. The table documents the prompt text alongside the top two [MASK] generated by the model and recorded results from ClinicalBERT. The last column indicates the prompt type. For medicine prediction, additional categories such as GPT-generated prompts or manually generated ones along with positive(p) and negative(n) should be included.

Task Category	Prompt Text	Predicted [MASK]	Type of Prompt
Medical Diagnosis Prediction	Given the patient's symptoms of [symptoms], the most likely diagnosis is [MASK] Given the patient's symptoms of [symptoms], it is unlikely that the diagnosis is [MASK]	As per the list of symptoms -	Positive Tone Negative Tone
Treatment Recommendation	Considering the patient's condition of [conditions] and history of [history], the recommended treatment would be [MASK] Considering the patient's condition of [condition] and history of [history], it would be inappropriate to recommend [MASK]	As per the list of conditions -	– –
Patient Triage and Prioritization	Given the urgency of the patient's symptoms [symptoms], prioritize this patient for immediate care. Given the patient's stable condition [symptoms], this patient can safely wait for regular medical attention.	– –	– –
Medical Image Interpretation	The features present in this image [image features] indicate the likelihood of [condition/disease] The features present in this image [image features] do not indicate the presence of [condition/disease].	– –	– –
Clinical Trial Matching	The patient's profile [patient profile details] matches the criteria for the clinical trial [trial name]. The patient's profile [patient profile details] does not match the criteria for the clinical trial [trial name].	– –	– –
Personalized Health Suggestions	Given the patient's lifestyle and health records [details], the following health improvements are suggested [MASK] Given the patient's lifestyle and health records [details], the following actions are not recommended [MASK]	– –	– –
Automated Medical Coding	The correct medical code for the described procedure [procedure details] is [MASK]: Assigning the medical code [code] for the described procedure [procedure details] would be incorrect.	– –	– –
Drug Interaction and Side Effects Analysis	Given the patient's current medication [medication], potential interactions and side effects could include [MASK]: Given the patient's current medication [medication], it is unlikely to cause the following interactions or side effects [MASK]:	– –	– –
Mental Health Analysis	The patient's responses [patient responses] suggest a high likelihood of [mental health issue]. The patient's responses [patient responses] do not indicate a significant risk of [mental health issue].	– –	– –
Genomics Data Interpretation	The presence of [genetic markers] in the patient's genomics data suggests a predisposition to [condition/disease]. The absence of [genetic markers] in the patient's genomics data reduces the likelihood of [condition/disease].	– –	 –
Medicine Prediction	Taking {medicine} has been [MASK] for your health Taking {medicine} has been not [MASK] for your health	Important, Appropriate Important, Sufficient	gpt_prompt1p gpt_prompt1n
	⋮	⋮	⋮
	Research studies have shown that {medicine} is [MASK] in reducing pain Research studies have shown that {medicine} is not [MASK] in reducing pain	Effective, Interested Interested, Effective	gpt_prompt6p gpt_prompt6n
	The medicine you are taking is [MASK] for you The medicine you are taking is not [MASK] for you	Good, Relevant Tested, Taking	manual_prompt1p manual_prompt1n
	⋮	⋮	⋮
	The medicine has [MASK] your health The medicine has not [MASK] your health	Changed, Helped Changed, Affected	manual_prompt4p manual_prompt4n

the pre-trained language model's predictions and enhancing its comprehension of certain healthcare tasks. Using a set of prompts instead of a single prompt in prompt engineering, particularly in healthcare applications, offers several advantages.

- Increased Model Robustness: A set of prompts can capture a wider range of linguistic variations and nuances, leading to more robust model performance. It reduces the model's dependency on a specific phrasing or structure of the prompt, making its predictions more reliable and less prone to errors due to unexpected input formats.
- Enhanced Context Understanding: Different prompts can provide various perspectives or highlight different aspects of the input data. This multi-perspective approach can lead to a more comprehensive understanding of the context, enabling the model to make more informed and accurate predictions.
- Reduced Prompt Bias: Relying on a single prompt can introduce bias, as the model's output may overly depend on that prompt's specific wording or structure. Using a diverse set of prompts mitigates this risk, ensuring that the peculiarities of a single prompt do not unduly influence the model's predictions.
- Improved Generalizability: A diverse set of prompts can cover a broader range of language styles and terminologies, making the model more adaptable to different datasets or domains. This is particularly important in healthcare, where terminology and presentation vary significantly across different specialties or institutions.
- Facilitation of Comparative Analysis: Using multiple prompts allows for a comparative analysis of the model's outputs. This can be particularly useful in uncertain tasks, as the model's responses to different prompts can be compared and contrasted to arrive at a more confident prediction or recommendation.
- Support for Error Analysis and Model Debugging: When the model's predictions based on different prompts are inconsistent, it can signal potential issues or areas for improvement in the model's understanding of the data. This can be valuable for error analysis and subsequent model refinement.

In conclusion, employing a set of prompts rather than a single prompt contributes to models' robustness, accuracy, and reliability, especially in complex, high-stakes fields like healthcare. It ensures that the model can handle a variety of inputs effectively and make well-informed predictions or recommendations.

5 EXPERIMENTS USING CONTEXT-AWARE PROMPT

The subsequent subsections present the experimental investigation of the proposed Prompt-Eng for medicine prediction conducted on the MIMIC-III's medical text corpus and MIMIC-CXR x-ray images.

5.1 MIMIC-III Experiments

To initiate the medicine prediction task, we first employed the traditional machine learning approach, transforming 500 random records from a subset of 36K samples prepared for experiments, including patient tabular and raw summary data, into tf-idf form, with the medicine names (list) as utilized as labels for prediction. The machine learning algorithm experiment is shown in Figure 2. This approach demonstrates good performance but achieves with significant resources and is computationally expensive, the minimum time recorded with M30 were 3150 seconds. To leverage the power of a pre-trained model and fine-tuning approach, further experiments were conducted using context-aware prompt pairs as recorded in Table. 1 and the prompt template shown in the gray rectangle box of the proposed architecture diagram. We conducted experiments on various ensemble measures to emphasize and validate that integrating prompts with pre-trained models boosts the

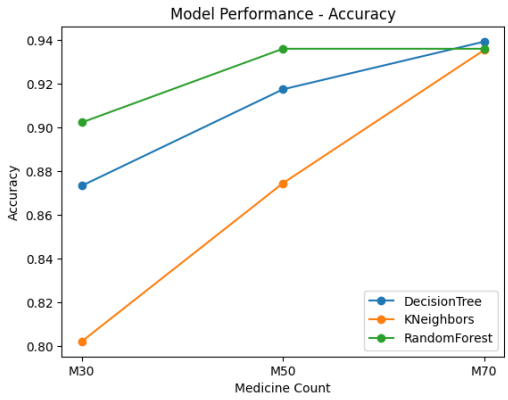


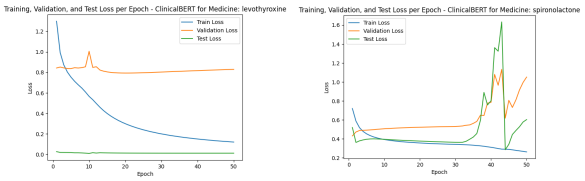
Figure 2: Accuracy Metric Comparative Analysis of Medicine Set using Traditional Machine Learning Algorithm

Table 2: Comparative Evaluation of Ensemble Measures for Medicine Prediction using Prompts with 500 records of Clinical Corpus (without fine-tuning). Blue highlighted metrics show the best result. In contrast, red shows the second optimal result.

Backbone Model	MedCount	EnsembleMeasure	Accuracy
BERT	M30	Mean Average	0.882
		Max Average	0.832
		Min Average	0.839
		Using Majority Voting	0.205
	M50	Mean Average	0.064
		Max Average	0.068
		Min Average	0.067
		Using Majority Voting	0.064
	M70	Mean Average	0.913
		Max Average	0.883
		Min Average	0.887
		Using Majority Voting	0.153
ClinicalBERT	M30	Mean Average	0.8332
		Max Average	0.882
		Min Average	0.593
		Using Majority Voting	0.350
	M50	Mean Average	0.869
		Max Average	0.914
		Min Average	0.635
		Using Majority Voting	0.399
	M70	Mean Average	**0.922**
		Max Average	**0.937**
		Min Average	0.680
		Using Majority Voting	0.446

results. This study evaluates several ensemble metrics (with different model configurations) to improve the accuracy of medicine prediction. Four aggregation procedures were utilized: Mean Average, Max Average, Min Average, and Majority Voting. The Mean Average determines the average prediction probability across all models, offering a fair depiction of the ensemble's overall prediction confidence. On the other hand, Max Average chooses the forecast with the highest confidence level among the models, while Min Average opts for the most cautious prediction. Majority Voting utilizes a voting system. Every metric provides distinct perspectives on the ensemble's ability to forecast, thoroughly comprehending the performance of medical prediction in different settings and model setups. The results are recorded in Table. 2 for the different medicine sets and prompts designed as depicted in Table. 1.

Furthermore, fine-tuned results are emphasized by showing the mean prediction using the fine-tuned ClinicalBERT model as recorded in the result in Table. 4. Compared to the traditional approach, the pre-trained model approach (with and without fine-tuning) took an average time from 2.5/epoch to 4.36/epoch, respectively. Notably, Prompt-Eng achieved results with significantly reduced computational time.



Levothyroxine **Spironolactone**

Figure 3: Training, validation, and testing loss plots using 50 epochs with Fine-tuned ClinicalBERT employed within Prompt-Eng

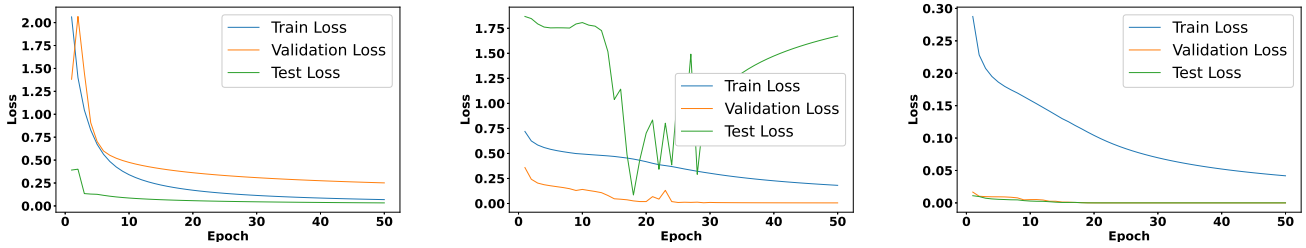
5.2 MIMIC-CXR Experiments

This section explores the comparative analysis of visual and textual prompt combinations designed and employed with VGG16 and VGG19 pre-trained models and our proposed Prompt-EngCNN. Overall, context-aware prompt pairs are emphasized throughout experimentation to improve model generalization and performance in healthcare applications.

To explore visual prompts applied to X-ray image analysis, we conducted experiments to enhance chest X-ray diagnoses' interpretability and predictive accuracy. Beginning with a self-designed simplistic model utilizing Convolutional Neural Networks (CNNs). The overall Prompt-EngCNN is designed to learn hierarchical representations of features from input images, gradually transforming them into high-level representations suitable for classification tasks, whether Pneumonia exists or not. We examined the efficacy of incorporating visual prompts to augment the model's predictive capabilities. In Table. 3 and Figure 6, We observed notable improvements in predictive performance by applying distinct visual prompts, such as highlighting specific regions of interest within the X-ray images. Then, we added text prompts, augmenting the model's interpretability and potentially enhancing its diagnostic accuracy. Experiments involving prompts tailored to emphasize critical features within the X-ray images yielded more precise predictions than the baseline model. These findings underscore the significance of incorporating visual prompts and text prompts as complementary tools in medical image analysis, offering potential avenues for enhancing diagnostic accuracy and facilitating clinical decision-making in healthcare settings. We tested a few Subject IDs from the enriched MIMIC-CXR dataset, and here in this study, we have shown various combinations of experiments as shown in Figure 6 and Figure 5. We used the constant (same) "Highlight region" related prompt for these experiments; thereby, the highlighted regions are the same, but the predicted score is different and depends on the backbone

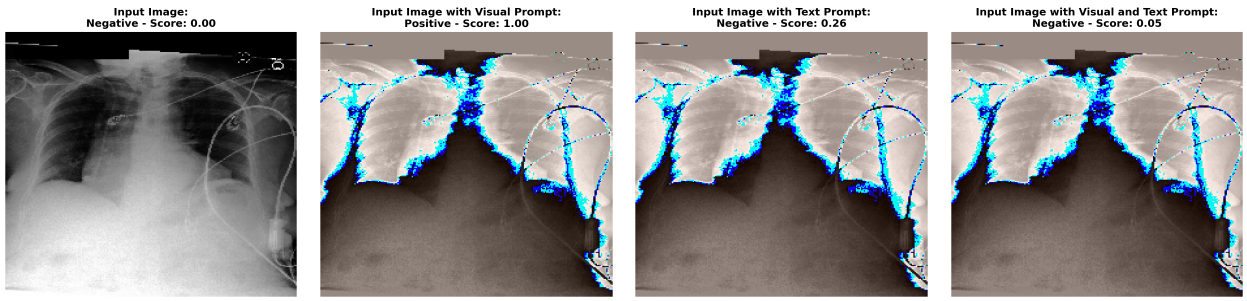
Table 3: Comparative Table Analysis of Different Prompt Engineering with Different Backbones. The table shows distinctive results, such as no constant prediction across different prompting techniques or backbones. These findings underscore the nuanced impact of prompt engineering on model performance, and further improvement and validation are suggested. Red highlighted shows the optimal result in the absence of the best. P=Positive Label and N=Negative

MIMIC-CXR SubjectID	Backbone	Actual Label	Initial prediction (No prompting)	Visual Prompt Prediction	Text Prompt Prediction	Visual+Text
s53254222	Prompt-EngCNN	N	N	N	N	P
	VGG16	-	P	P	N	P
	VGG19	-	P	P	P	P
s59258773	Prompt-EngCNN	Doubtful(unlabeled)	P	P	N	P
	VGG16	-	P	P	N	P
	VGG19	-	P	P	P	P
s56466802	Prompt-EngCNN	Doubtful(unlabeled)	N	P	N	N
	VGG16	-	P	P	P	N
	VGG19	-	P	P	N	P



Predictions with Manual Prompt Set **Predictions with GPT Suggest Prompt Set** **Predictions with all combined Prompt Set**

Figure 4: Comparative analysis of training, validation, and testing loss plots with different sets of prompts. The loss plots suggest that manually crafted or carefully designed prompts contribute to the model's learning process, improving performance and convergence during training and validation.



inputImageCNN **visualPromptCNN** **TextPromptCNN** **visualTextPromptCNN**

Figure 5: Comparative Analysis of Prompt Engineering using Different Backbones - MIMIC-CXR SubjectID = s56466802. Visual Prompt is "Highlight a region visual_prompt[:, 100:150, 100:150, :] = 1", Text Prompt is "This is an X-ray image of the chest"

Table 4: Comparative Results for Ensemble Set of Prompts (with fine-tuning). Blue highlighted shows best results achieved during training of fine-tuned ClinicalBERT model.

MedCount	Accuracy
M30	**0.991**
M50	0.761
M70	0.657

model's capability. Other visual prompts for an image include but are not limited to "checkerboard pattern", 'highlight top boundary," "highlight left boundary," "Gaussian filter," "Gradient filter," and "random noise." all such prompts are specific to task such as our task is interpretation of X-Ray image. To further validate the remaining results, we prepared a table where all three Subject ID experiments are recorded.

6 DISCUSSION

By harnessing the inherent knowledge of language models, our approach eliminates the need for extensive pre-training on medical datasets, thus saving computational resources and time. We extensively explore prompt techniques to enhance the accuracy of

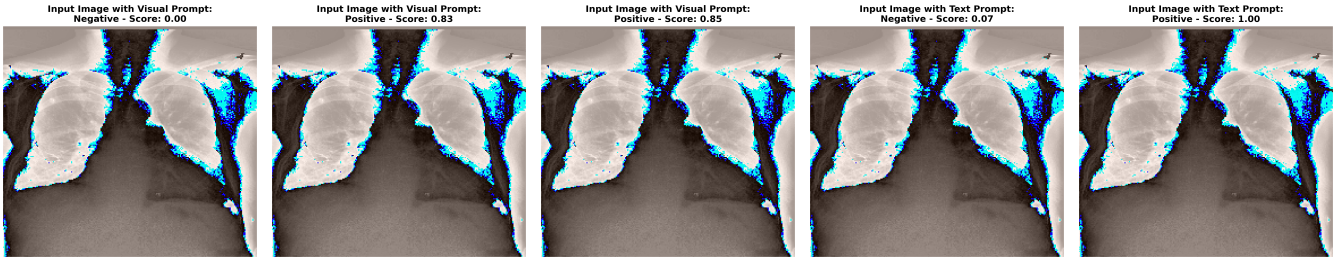


Figure 6: Comparative Analysis of Different Prompt Engineering with Different Backbones - MIMIC-CXR SubjectID = s53254222. Visual Prompt is "Highlight a region visual_prompt[:, 100:150, 100:150, :] = 1", Text Prompt is "This is an X-ray image of the chest, find pneumonia"

medicine prediction. Through a series of experiments on various medicine cohorts, including 30, 50, and 70 medicines, we evaluate the performance of our model. The results demonstrate the effectiveness of prompt techniques in achieving accurate predictions, even without traditional training methodologies. Our experiments on the medical diagnostic category show promise in reducing the dependency on labeled medical data, making it a valuable tool for personalized healthcare and treatment recommendation systems.

Additionally, one notable aspect of our research is the integration of visual prompts on X-ray images. This technique is promising for enhancing our model's interpretability and predictive power. By incorporating visual and textual prompts, we provide the model with additional contextual information, enabling it to make more informed predictions. These interactions of visual and textual prompts represent a novel approach in healthcare prediction systems, offering new avenues for improving diagnostic accuracy and treatment recommendation.

Our research advances healthcare prediction by providing a training-free model that leverages prompt techniques for accurate and efficient predictions. The findings open new possibilities for enhancing healthcare decision-making and improving patient outcomes. They have the potential to revolutionize personalized healthcare and treatment recommendation systems.

7 LIMITATIONS

Below, we acknowledge a few limitations that warrant attention in future work.

- ***Scope:*** The primary emphasis of the work is the use of prompt engineering approaches in the healthcare field.
- ***Experimental Generalizability:*** Although we performed various experiments, we were limited because our scope of the study was only to emphasize the importance of prompt engineering. Further detailed experiments are needed to justify the generalizability.
- ***Comprehensive Training and Fine-tuning should be conducted:*** Since our experiments only involve fine-tuning of ClinicalBERT as per Algorithm. 3 for medicine prediction based on prompts as presented in Table. 1. However, our models underwent minimal training for other categories, such as X-ray interpretation by individual screening with various backbone architectures. Further training and fine-tuning with larger and more diverse datasets is essential to validate the effectiveness and robustness of employing prompts with different backbone architectures.
- ***Evaluation metrics:*** The study suggests validating other metrics would benefit the model's effectiveness, as we only present accuracy, which can not guarantee the comprehensive evaluation of the model.
- ***Computational Resources:*** This study does not focus on evaluating the utilization of computational resources. Instead, we aimed to assess the effectiveness of different prompt sets for improving medical diagnostic and medical prediction; we recommend conducting such experiments to evaluate the computational efficiency.

8 CONCLUSION

This study presents Prompt-Eng, a novel framework for prompt engineering in healthcare applications. Our study concludes that leveraging pre-trained language models and prompt engineering proves effective across diverse healthcare tasks. Providing instructive prompts enables language models to extract insights, make accurate predictions, and assist in clinical decision-making. Our exploration of prompt engineering in healthcare highlights its potential to revolutionize healthcare practice. By designing precise prompts for specific tasks and scenarios, we enhance the interpretability and reliability of model predictions. Prompt pairs, including positive and negative aspects, further enhance model generalization and performance across diverse tasks.

Furthermore, our investigation emphasizes the importance of prompt optimization strategies and context-aware prompts for multimodal data analysis, and integrating prompt engineering would also enhance the robustness of the model. In conclusion, Prompt-Eng offers a promising avenue for advancing healthcare applications through prompt-guided language models. By harnessing the power of prompt engineering, we can unlock opportunities for leveraging large-scale clinical data, improving diagnostic accuracy, optimizing treatment recommendations, and ultimately enhancing patient outcomes. As prompt engineering continues to evolve, we anticipate further innovations that will reshape healthcare delivery and usher in a new era of precision medicine.

RESOURCE ACKNOWLEDGEMENT

The dataset (subset) and relevant code files will be available at the GitHub repository [3] after formal publication in the Companion Proceeding of the ACM Web Conference 2024.

[3] https://github.com/awaisahmednuces/PrompEng-theWebConferenceWorkshopPaperRep

REFERENCES

[1] Shuroug A Alowais, Sahar S Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I Alshaya, Sumaya N Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A Badreldin, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, 23(1):689, 2023.

[2] Awais Ahmed, Rui Xi, Mengshu Hou, Syed Attique Shah, and Sufian Hameed. Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts. *IEEE Access*, 2023.

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

[4] Awais Ahmed, Xiaoyang Zeng, Rui Xi, Mengshu Hou, and Syed Attique Shah. Med-prompt: A novel prompt engineering framework for medicine prediction on free-text clinical notes. *Journal of King Saud University-Computer and Information Sciences*, page 101933, 2024.

[5] Jianing Qiu, Lin Li, Jiankai Sun, Jiachuan Peng, Peilun Shi, Ruiyang Zhang, Yinzhao Dong, Kyle Lam, Frank P-W Lo, Bo Xiao, et al. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[6] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

[7] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113, 2022.

[8] Siru Liu, Aileen P Wright, Barron L Patterson, Jonathan P Wanderer, Robert W Turer, Scott D Nelson, Allison B McCoy, Dean F Sittig, and Adam Wright. Using ai-generated suggestions from chatgpt to optimize clinical decision support. *Journal of the American Medical Informatics Association*, 30(7):1237–1245, 2023.

[9] Haoxin Liu, Wenli Zhang, Jiaheng Xie, Buomsoo Kim, Zhu Zhang, and Yidong Chai. Few-shot learning for chronic disease management: Leveraging large language models and multi-prompt engineering with medical knowledge injection. *arXiv preprint arXiv:2401.12988*, 2024.

[10] Majid Afshar, Yanjun Gao, Graham Wills, Jason Wang, Matthew M Churpek, Christa J Westenberger, David T Kunstman, Joel E Gordon, Frank J Liao, and Brian W Patterson. Prompt engineering gpt-4 to answer patient inquiries: A real-time implementation in the electronic health record across provider clinics. *medRxiv*, pages 2024–01, 2024.

[11] Daphne van Zandvoort, Laura Wiersema, Tom Huibers, Sandra van Dulmen, and Sjaak Brinkkemper. Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting. *arXiv preprint arXiv:2311.13274*, 2023.

[12] Niall Taylor, Yi Zhang, Dan W Joyce, Ziming Gao, Andrey Kormilitzin, and Alejo Nevado-Holgado. Clinical prompt learning with frozen language models. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[13] Fatemeh Nazary, Yashar Deldjoo, and Tommaso Di Noia. Chatgpt-healthprompt. harnessing the power of xai in prompt-based healthcare decision support using chatgpt. In *European Conference on Artificial Intelligence*, pages 382–397. Springer, 2023.

[14] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *Meta-Radiology*, page 100047, 2023.

[15] Sabit Ekin. Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices. *Authorea Preprints*, 2023.

[16] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*, 2023.

[17] Bertalan Meskó. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of Medical Internet Research*, 25:e50638, 2023.

[18] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[19] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

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

[21] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020.