Prompt Engineering for Healthcare: Methodologies and Applications

Jiaqi Wang [*][1], Enze Shi [*][1], Sigang Yu[1], Zihao Wu[3], Chong Ma[2], Haixing Dai[3], Qiushi Yang[10], Yanqing Kang[1], Jinru Wu[1], Huawen Hu[1], Chenxi Yue[1], Haiyang Zhang[1], Yiheng Liu[11], Xiang Li[4], Bao Ge[11], Dajiang Zhu[5], Yixuan Yuan[9], Dinggang Shen[6,7,8], Tianming Liu[3], and Shu Zhang [†][1]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
[2] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[3] School of Computing, The University of Georgia, Athens 30602, USA
[4] Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston 02115, USA
[5] Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington 76019, USA
[6] School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China
[7] Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China
[8] Shanghai Clinical Research and Trial Center, Shanghai, 201210, China
[9] Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong 999077, China
[10] Department of Electronic Engineering, City University of Hong Kong, Hong Kong 999077, China
[11] School of Physics and Information Technology, Shaanxi Normal University, Xi'an 710119 China

Abstract

This review will introduce the latest advances in prompt engineering in the field of natural language processing (NLP) for the medical domain. First, we will provide a brief overview of the development of prompt engineering and emphasize its significant contributions to healthcare NLP applications such as question-answering systems, text summarization, and machine translation. With the continuous improvement of general large language models, the importance of prompt engineering in the healthcare domain is becoming increasingly prominent. The aim of this article is to provide useful resources and bridges for healthcare NLP researchers to

---
[*]Co-first author
[†]Corresponding author: shu.zhang@nwpu.edu.cn

better explore the application of prompt engineering in this field. We hope that this review can provide new ideas and inspire ample possibilities for research and application in medical NLP.

1   Introduction

Prompt engineering has emerged as a cutting-edge approach in the field of NLP, providing a more efficient and cost-effective means of using large language models (LLM). This innovative paradigm is rooted in the development of LLMs, which have transformed our understanding of natural language [1, 2].

The history of language models can be traced back to the 1950s, but it is until the introduction of the BERT and GPT models in 2018 that LLMs became the mainstream [3]. These models utilize complex algorithms and massive amounts of training data to understand natural language in ways that are previously unimaginable. Their ability to learn patterns and structures in language has proven invaluable in a wide range of language-related tasks [4, 5].

However, fine-tuning these pre-trained models can be a costly and time-consuming process, requiring significant amounts of annotated data and computing resources. To address this challenge, researchers have turned to prompts as a means of guiding model learning [2, 6, 7]. Prompt learning is a novel paradigm in NLP that enables language models to perform few or even zero-shot learning, adapting to new scenarios with minimal labeled data [1, 2, 8]. This approach is rooted in language modeling, directly modeling the probability of text. The key to prompt engineering is designing prompts for downstream tasks, which guide the pre-trained model to perform the desired task [9].

Prompt learning can be broken down into five key steps [1]. First, researchers must choose an appropriate pre-training model. Next, they must design prompts for downstream tasks, which can be tailored to the specific requirements of each task, and this step is called the prompt engineering process. The third step involves designing responses based on the task at hand, allowing the model to produce the desired output. The fourth step is to expand the paradigm to further improve results or adaptability methods. Finally, researchers must design training strategies that enable the model to learn efficiently and effectively. Prompt engineering is thus a critical step in the prompt-based approach, which involves designing effective prompts to guide the pre-trained language model in downstream tasks. While prompt engineering plays a crucial role in the success of the overall approach, studies have shown that the design of prompts can significantly impact the performance of the model on downstream tasks. A well-designed prompt should provide clear guidance to the model and facilitate effective task completion. Moreover, optimizing prompt parameters is also an important aspect of prompt engineering, as it can lead to improved model performance. The subsequent steps in prompt-based methods, such as designing answers, expanding the paradigm, and adjusting training strategies, require task-specific prompt design and optimization.

Overall, prompt engineering is a promising new approach in NLP that has

the potential to revolutionize the field. Its focus on using prompts to guide model learning provides a more cost-effective and efficient means of training large language models, making it an increasingly important area of research in the years to come [10].

Figure 1: A wordcloud is employed to illustrate the research hotspots, focal points, and directions in prompt engineering for NLP in the medical domain, which reflects the key vocabulary and highlights the significant themes of this literature review.

1.1 Scope and Focus of the Review

In recent years, the healthcare industry has received increasing attention as it is closely related to each individual. However, there is still a significant gap between the shortage of medical resources and people's needs [11, 12]. NLP technology can handle massive amounts of data from various medical literature and medical records, thus helping doctors and patients better understand and manage diseases, which is of great significance in improving medical standards and providing better health security [13].

However, traditional machine learning and deep learning methods cannot solve NLP tasks in the medical field very well [3]. The complex and diverse professional terminology in the medical field, the wide range of professional knowledge, and ethical issues related to patient privacy have all become diffi-cult problems in the development of the healthcare industry [14]. The emergence of LLMs and prompt-based methods provides a new solution for NLP tasks in the medical field. As depicted in Figure 1, the word cloud illustrates the re-search hotspots and focal points in NLP for the medical domain, highlighting the significant themes of this literature review in this field. Through the pow-erful contextual learning ability of LLMs, useful information can be effectively

obtained from a large number of medical literature and cases. By designing specific prompts, domain-specific knowledge and task-specific information can be introduced into the pre-trained model, resulting in better performance and gaining extensive attention [15].

In the research of prompt engineering, designing appropriate prompts is an important issue. Currently, researchers are exploring various types of prompts, including manual prompts [14–19] and automated prompts [6, 20–22], as well as various methods of prompt construction to help solve tasks in the medical field.
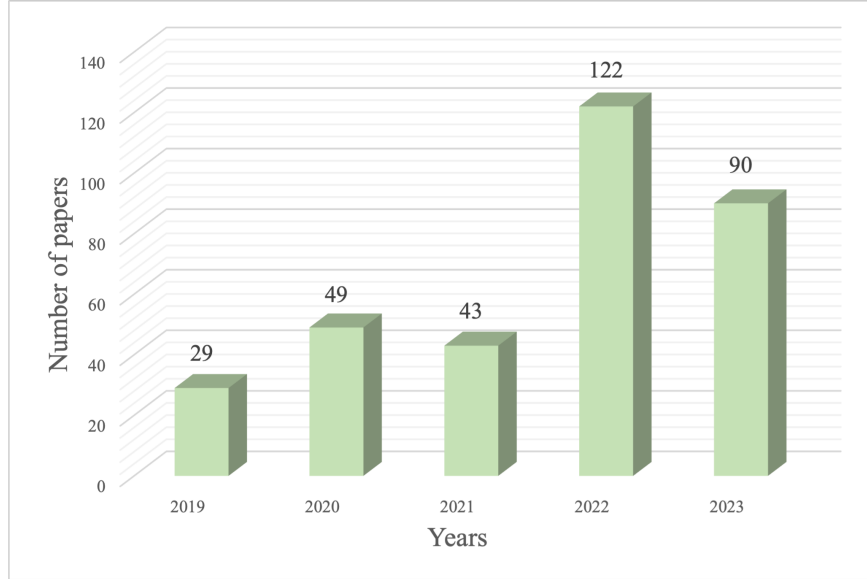
Figure 2: A graphical representation is utilized to depict the number of research papers on prompt engineering for NLP in the medical domain, published from 2019 to April 6, 2023, revealing the trend and growth of this field over time.

1.2 Literature Search Process

The importance of NLP tasks in the medical field cannot be overstated. With the increasing volume of medical data, there is a growing need for effective NLP techniques to improve the quality and efficiency of medical services. In this regard, prompt engineering has emerged as a promising approach to guiding model generation by providing targeted prompt information.

To investigate the application of prompt engineering in the medical field, we conduct a comprehensive review of literature from 2019 to 2023 by searching the keywords "prompt" and "medical NLP" on arxiv. As shown in Figure 2, a total of 333 papers related to the prompt are identified, and ChatGPT is used to screen the abstracts for their relevance to the medical field, resulting in the

selection of 140 relevant papers for further review. The dynamic graph in the following link[1] shows the daily publication rate of related papers.

These papers mainly focus on the design and application of prompt engineering in NLP tasks in the medical field, with the goal of providing suitable prompt design methods for different medical tasks. Specifically, the studies explore how to choose and design prompt elements, how to use prompts to guide a model generation of text that meets medical requirements, and how to evaluate the impact of different prompt designs on model performance.
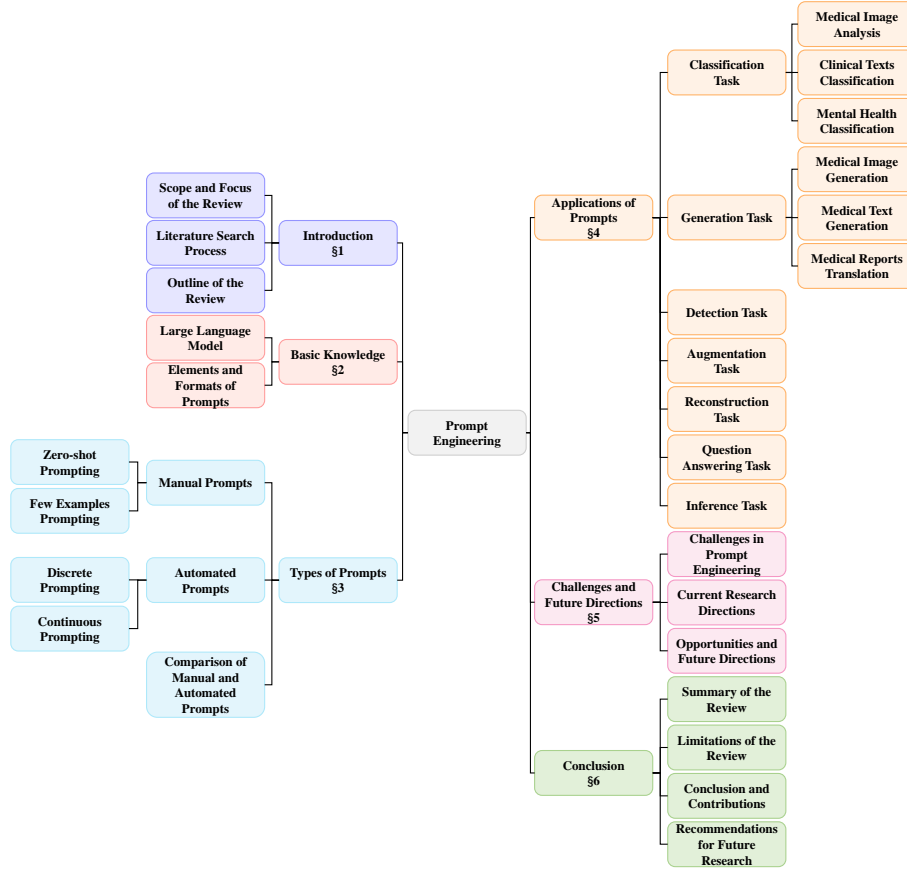
Figure 3: The overview of this literature review provides a comprehensive and systematic summary of the current state of research on prompt engineering for NLP in the medical domain.

---

[1] https://braininspiredai.github.io/prompt_trend_in_medical/

1.3 Outline of the Review

This review article provides an in-depth overview of Prompt Engineering, a rapidly growing field of research focusing on enhancing medical applications through the development of effective prompts. As illustrated in Figure 3, we present the overview of this review article with each section organized as follows:

The introduction of the article discusses the background and significance of Prompt Engineering and outlines the scope and focus of the review. It describes the literature search process followed and provides an outline of the article.

Section II presents the basics of Prompt Engineering, including common LLMs, and the elements and components of prompts.

Section III discusses the different types of prompts available in the literature, with a focus on manual prompts such as zero-shot and few-shot prompting. The section uses medical literature to highlight specific details. Automated prompts, including discrete and continuous prompting, are also discussed, and the section compares manual and automated prompting techniques.

In Section IV, the various applications of prompts in medical fields are covered in detail. These applications include classification, generation, detection, augmentation, reconstruction, question answering, prediction, and inference tasks, and the section provides the latest medical examples to illustrate the practical use of prompts.

Section V outlines the challenges and future directions of prompt engineering in medical applications. The section highlights the current research directions in the field and provides insight into the opportunities for future research.

Finally, Section VI concludes the review by summarizing the key findings and contributions of the article. The section also discusses the limitations of the review and provides recommendations for future research in the field of prompt engineering in medical applications.

Overall, the review provides insights into the latest developments in prompt engineering for medical NLP tasks and highlights the importance of prompt engineering in improving the accuracy and effectiveness of medical services. It also underscores the need for continued research to explore the potential of prompt engineering in the medical field and to identify effective strategies for its implementation. Ultimately, prompt engineering has the potential to transform the way medical services are delivered and to enhance the quality of care for patients.

2 Basic Knowledge

In prompt engineering, LLMs are of great importance as they enable the design of appropriate prompts for specific tasks and can generate high-quality results even with limited training data [16]. This section will delve into the latest advances in large models and provide a framework for prompt design.

2.1 Large Language Models

In this section, we have explored the prevalent LLMs in prompt engineering within the medical field, which are trained on large amounts of text data and are capable of generating high-quality, contextually relevant, and coherent text in various NLP tasks. Among them, prompt-based techniques provide researchers with rich natural language patterns and structures and have become an important tool in prompt engineering.

Prompt engineering is achieved through the design of prompts, which enables the model to generate diversified text based on different contextual environments and can be optimized and customized for different tasks and application scenarios [23]. In the medical field, the application of these models has provided significant help and inspiration for medical research and clinical practice.

BERT (Bidirectional Encoder Representations from Transformers) [24] is a pre-trained model developed by the Google team in 2018 that has bidirectional encoder representations using the transformer architecture. Its purpose is to pre-train bidirectional representations by jointly adjusting the context in all layers, and by overcoming the limitations of previous unidirectional language models through masked language model. In addition, BERT introduces the next sentence prediction task, which can be used in conjunction with masked language model to pre-train text pair representations [25]. The training objective of BERT is highly meaningful, as it demonstrates the importance of bidirectional pre-training for language representations. Compared to previous unidirectional language models, BERT's pre-training can better capture semantic information in the context [24].

BERT is the first fine-tuning based representation model, which has surpassed many task-specific architectures in 11 NLP tasks, setting new performance records [24]. This achievement demonstrates the enormous potential of pre-trained models in NLP tasks, and provides important references for the development of subsequent NLP models.

The Text-to-Text Transfer Transformer (T5) [4] is an LLM developed by Google in 2019. The fundamental idea behind T5 is to treat every NLP task as a "text-to-text" problem. The training objective of T5 is to develop it into a universal knowledge representation model, enabling the model to better understand and process text. T5 has achieved state-of-the-art results in multiple NLP tasks, such as text classification, machine translation, text summarization, and question answering, among others. In these tasks, the text is taken as input, and new text is generated as output to solve specific problems. The advantage of this approach is that by treating all NLP tasks as "text-to-text" problems, the T5 model can learn how to map input text to output text, thereby enhancing its understanding and processing of text.

ChatGPT is a series of generative pre-trained transformer models developed by the OpenAI team. GPT-1 [26], which is released in June 2018, is trained on a large-scale text corpus with the aim of generating coherent and natural text. Subsequently, GPT-2 [27] is released in February 2019, which is built on the foundation of GPT-1 with a larger model size and more training data, achieving

7

outstanding performance on a variety of NLP tasks. In June 2020, GPT-3 [28] is introduced, utilizing an even larger model size and more extensive training data than GPT-2, resulting in further improvements in model performance. GPT-3.5 then further optimizes the efficiency and performance of GPT-3.

To enhance the performance of the model, the OpenAI team uses supervised and reinforcement learning to fine-tune GPT-3.5, with human intervention playing a key role in enhancing machine learning capabilities [29, 30]. The optimization of training methods and increased training data have led to GPT-4 being able to generate text that is similar to that produced by humans, representing a significant breakthrough in the field of artificial intelligence.

In the future, LLMs will continue to evolve and improve, and we can expect their application in the medical field to bring more value and contributions to human health.

2.2 Elements and Formats of Prompts

Prompt engineering refers to the process of creating a prompt function that guides a model to generate effective performance in downstream tasks. This process typically involves two steps. First, a template is applied, which is a text string that contains two slots: an input slot [X] for the input text x and an answer slot [Z] for the intermediately generated answer text z that will later be mapped into the final output y. Then, the input slot [X] is filled with the input text x [1].
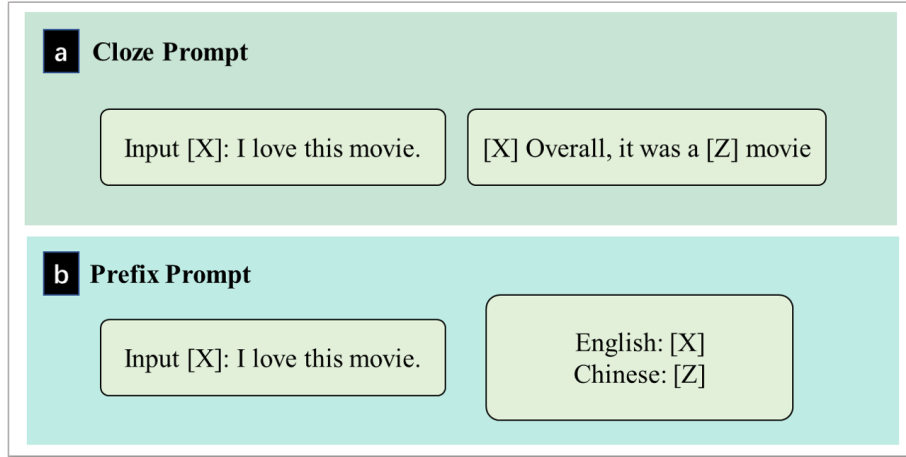

Figure 4: This study distinguishes between two types of prompts, namely cloze prompts (illustrated in a) and prefix prompts (illustrated in b), where slot [x] represents the input text and slot [z] represents the answer text.

In NLP, prompts are usually classified into two forms: cloze prompts [16, 31] and prefix prompts [6, 22]. Cloze prompts refer to a slot that needs to be filled

in the middle of the template text. For example, in sentiment analysis, when X = "I love this movie," the template may take the form of " [X]. Overall, it was a [Z] movie." Then, X' becomes "I love this movie. Overall it was a [Z] movie." This is an example of cloze prompts as shown in Figure 4(a). Prefix prompts, on the other hand, refer to the input text being entirely before the generated answer text Z. As shown in Figure 4(b), "English: [X] Chinese:[Z]".

It is important to note that in prompt engineering, prompts are not only in the form of natural language but can also generate continuous vector embeddings [22]. This part will be discussed in detail in section three. By utilizing prompt engineering techniques, model performance, and efficiency can be significantly improved, providing better support and assistance for downstream tasks.

3   Types of Prompts

In the realm of scientific research, prompts play a crucial role in guiding large models to produce accurate outputs. Prompts are broadly categorized into two types: manual prompts and automated prompts [1]. Manual prompts are meticulously crafted by human experts to provide explicit instructions to the model on what type of data to focus on and how to approach the task at hand. These prompts are particularly effective when the input data is well-defined and the output must conform to a specific structure or format.

However, manual prompts have certain limitations that cannot be overlooked. Creating effective manual prompts requires significant expertise and time, and even minor changes to the prompts can have a significant impact on model predictions [14–19]. This is especially true for complex tasks where providing clear and effective prompts can be a challenging endeavor. To address these limitations, researchers have developed various automated methods to design prompts.

Automated prompts have gained popularity due to their efficiency and adaptability. These prompts are generated using various algorithms and techniques, eliminating the need for human intervention [6, 20–22]. Discrete prompts and continuous prompts are two common types of automated prompts. Discrete prompts rely on predefined categories to generate responses, while continuous prompts consider the current conversation context to produce accurate outputs [1, 32–35]. Additionally, there are static prompts and dynamic prompts that vary in their consideration of historical context [36, 37].

3.1   Manual Prompts

Manual prompts can be used to guide models towards producing the desired results for specific tasks. These prompts are often created based on human expertise, and are commonly used in LLMs to improve their performance on downstream tasks. For instance, the LAMA dataset introduced hand-crafted cloze prompts to explore the knowledge of large models [16]. Similarly, prefix

prompts can be manually created to facilitate common sense reasoning tasks. In chat-oriented models like ChatGPT, zero-shot prompting is used to guide the models towards producing ideal results for downstream tasks. Additionally, it is common to include prompt examples that match the level of task complexity, to further guide models in producing accurate outputs. In this section, we mainly discuss the few-shot prompting that is manually created through instruction with examples selected by human experience, rather than being automatically selected. As AI technology rapidly advances, automated few-shot prompting techniques that select examples automatically are also emerging, which will be discussed in section 3.2.

3.1.1 Zero-shot Prompting

The superior performance of LLMs in a wide range of NLP tasks has garnered significant attention due to their remarkable ability to learn from context, making them an attractive tool for various research problems. This emerging capability is often enhanced through prompt engineering to improve the performance of LLMs in downstream tasks. Consequently, researchers have proposed various methods for designing prompts to guide the models' applications in different research questions [38].

Recently, the development of LLMs such as GPT-3 and chatGPT has made the use of prompt-based methods increasingly popular for various downstream tasks. Zero-shot prompting, in particular, has shown great promise, where providing a well-designed prompt alone without corresponding examples can lead to outstanding results. Some studies have suggested that leveraging the powerful performance of GPT can help us complete downstream tasks such as extraction and recognition, and even outperform some full-shot models on certain datasets [17]. In the medical field, the impressive achievements of using LLMs to exploit their strong contextual capabilities have also been observed. For instance, DeID-GPT proposes the use of high-quality prompts to preserve privacy and summarize key information in medical data on ChatGPT and GPT-4, achieving better results compared to several baseline methods. The study provides a foundation for future research on applying medical data to LLMs [14]. ChatAug proposes a prompt-based medical data augmentation method on Chat-GPT, which outperforms other popular methods. The study highlights the importance of domain knowledge in generating correct augmented data and suggests a fine-tuning approach for future research [15]. Recent work has investigated the feasibility of using manual prompts to guide downstream tasks on ChatGPT and has shown that ChatGPT's performance on translation tasks can be significantly improved with clear and accurate prompts [18]. Similarly, research on ChatGPT's zero-shot prompting has also yielded similar conclusions [19]. The study presents HealthPrompt, a prompt-based clinical NLP framework that explores different types of prompt templates, including prefix prompts and cloze prompts, to improve zero-shot learning performance on clinical text classification tasks without requiring additional training data [32]. These findings highlight the potential of ChatGPT to enhance NLP tasks with

effective prompt design, providing valuable insights for future research in this area.

Overall, the success of prompt-based methods has shown their great potential for further advancements, and their use is likely to continue to grow in the future.

3.1.2 Few-shot Prompting

While zero-shot prompting has shown impressive performance in many tasks, it still has some limitations. First, it heavily relies on the performance of pre-trained models. Secondly, due to the flexibility of zero-shot prompting, its outputs may not always be accurate. In such cases, providing a small number of prompt examples as guidance for the model to achieve better performance has become a useful approach. Few-shot prompting can serve as clear and explicit prompt inputs to guide the model towards desired outputs. For instance, research has been conducted to measure the baseline performance of GPT-4 on medical multiple-choice questions (MCQs) using only a few prompt samples, without the need for complex methods such as chain-of-thoug [39].

These prompt methods rely on the contextual emergence ability of LLMs, which has demonstrated impressive performance on ChatGPT/GPT-4 [40]. The manual prompt-based approach has shown typical few-shot prompting capabilities in the medical text translation, text augmentation, generation, summarization, and other tasks, with significant improvements over baseline models on public datasets [19, 41].

3.2 Automated Prompts

The effectiveness of human-designed prompts depends on the prompt designer's expertise in selecting relevant information and constructing prompts in a way that is understandable to the model. Human prompts are particularly effective for tasks with well-defined input data and structured output requirements. However, designing effective human prompts requires a significant amount of time and domain-specific knowledge, making their use in more complex tasks challenging. Therefore, researchers are exploring automated prompt design methods to address these challenges and improve the efficiency and adaptability of prompt-based approaches.

3.2.1 Discrete Prompting

Discrete prompts refer to automatically searching for templates in a discrete space, typically corresponding to natural language phrases, to provide the model with guidance for generating the desired output. This approach can help us design prompts more easily and improve the efficiency and adaptability of the model [1]. Common discrete prompt construction methods include prompt mining, prompt paraphrasing, prompt generation, and prompt scoring [42].

Prompt mining is a method of automatically discovering templates or prompts by scraping a large text corpus for frequent middle words or dependency paths

between a set of input-output pairs. These prompts can be used for various natural language processing tasks such as language modeling, text classification, and question answering. The research proposes a method for prompt-based learning that automatically identifies and adds prompt phrases to a prompt template for fine-grained detection of Alzheimer's disease. The method enhances detection accuracy through the use of additional prompts [43].

Prompt paraphrasing involves generating a set of candidate prompts from an existing seed prompt using methods such as round-trip translation or phrase replacement and selecting the one that achieves the highest training accuracy for a target task. This approach can be optimized using a neural prompt rewriter and can be done on an individual input basis. For instance, the article proposes a simple yet effective prompt method based on paraphrasing to assist pre-trained models in learning rare biomedical vocabulary, resulting in improved performance in biomedical applications [44–46]. The STREAM framework leverages prompt-based language models to generate task-specific logical rules for named entity tagging, reducing the need for human labor. By utilizing existing prompt templates and continue teaching the language model, STREAM achieves higher accuracy and efficiency in the tagging process [47].

Prompt generation is the task of generating prompts using natural language generation models. The study proposes MEDIMP, a framework that generates medical prompts using automatic textual data augmentations from LLMs, and demonstrates its superiority in producing high-quality prompts for medical applications [33]. The paper proposes a method that utilizes prompts filled with bounding box annotations to generate descriptions containing extensive hints and context for instance recognition and localization. The knowledge from language is then distilled into the detection model via maximizing cross-modal mutual information in both image and object-level [34]. This approach involves training prompts from existing data and using these prompts to generate new data. This newly generated data can then be used to define additional prompts, ultimately providing downstream tasks with more high-quality feature representations [35]. The study introduces a prompt construction module that leverages medical language templates to enable pre-trained language models to extract contextual information from tabular electronic health records and generate more comprehensive cell embeddings, which can be used by a pre-trained sentence encoder to generate sentence embeddings as cell representations [48].

Prompt scoring involves using language models to score filled prompts for a given task and selecting the one with the highest probability. This can result in a custom template for each individual input, as in the case of knowledge base completion. For instance, ImpressionGPT has proposed dynamic prompts, which select the top-k examples with the highest similarity for each input. Based on the probability and threshold of the k-selected examples, the prompt for each input is determined. This approach, which utilizes few-shot learning, has shown better performance than some full-shot learning methods [49].

The utilization of various prompt types can greatly benefit language models in their performance of a diverse range of tasks, such as natural language understanding, text generation, machine translation, and question-answering.

Each type of prompt, whether generated through prompt mining, prompt paraphrasing, prompt generation, or prompt scoring can be used independently or in combination to optimize the performance of the language model for a particular task. This underscores the significance of prompt-based approaches in advancing the capabilities of NLP systems.

3.2.2 Continuous Prompting

In recent years, the development of LLMs has made prompt-based methods increasingly popular in various downstream tasks. In this context, continuous prompting (e.g., soft prompting) has emerged as a new type of prompt and has gained widespread attention from researchers. Unlike discrete prompts, continuous prompts can be operated in the embedding space and are no longer limited to text-readable types. Additionally, continuous prompts can optimize parameters through tuning on downstream task training data, which relaxes the constraints on prompts and improves the efficiency of LLMs in task execution. Meanwhile, Continuous Prompting has been shown to be an effective alternative to standard model fine-tuning in cases where the data is highly imbalanced [50]. Discrete prompts are composed of discrete words or phrases, typically in human-interpretable natural language, to guide the model in performing downstream tasks. On the other hand, continuous prompts directly prompt the model in the embedding space, without constraints on natural language, and allow the prompts to have their own parameters that can be fine-tuned using training data.

In the medical field, the advantages of continuous prompting have been further demonstrated. Liang et al.(2022) use continuous prompts to introduce prompting into the model architecture as the only trainable parameter to guide model output, proposing PromptFuse and BlindPrompt as methods for aligning different modalities in a modular and parameter-efficient manner. These methods require only a few trainable parameters and perform comparably to several multimodal fusion methods in low-resource scenarios. The high modularity property of prompting allows for the flexible addition of modalities at low cost, as it avoids the need to fine-tune large pre-trained models [36]. PrefixMol is a generative model that utilizes a learnable prompt token as prefix embedding to guide the model to generate an output that meets the desired criteria. This prompt token encodes a set of learnable features and contextual information about the targeted pocket's circumstances and various properties. This approach offers significant advantages over traditional generative models as it allows for more precise and efficient control over the output [37]. Wang et al.(2022) propose an adaptive PromptNet for glioma grading that utilizes only non-enhanced MRI data and receives constraints from features of contrast-enhanced MR data during training through a designed prompt loss. In the paper, enhanced features are used as prompts to guide the feature extraction of PromptNet, with an adaptive strategy designed to dynamically weight the prompt loss in a sample-based manner, thus achieving competitive glioma grading performance on NE-MRI data [51]. CPP utilizes task-specific prompts opti-

mized with a contrastive prototypical loss to avoid semantic drift and prototype interference, using prompts as learnable tokens to train task-specific information for each task. It also introduces a multi-centroid prototype strategy to improve prototype representativeness. CPP outperforms existing methods under a light memory budget and improves class separation [52].

Overall, continuous prompting is a powerful tool for improving the performance of automated systems. It helps these systems to learn and adapt more quickly, and to perform more accurately in a wide range of applications. This technique can be used in combination with other types of prompts, such as manual prompts and discrete prompts, to create a more comprehensive and effective learning environment, which is also used in multi-modal tasks such as image and speech recognition to improve the accuracy and speed of the systems. Additionally, continuous prompts can be tailored to the specific needs of the task, which makes it highly adaptable and versatile.

3.3   Comparison of Manual and Automated Prompts

Manual prompts refer to human-crafted prompts, while automated prompts are generated by algorithms or automated methods. The primary difference between manual and automated prompts is the level of human involvement in the prompt construction process. Manual prompts are carefully designed by human experts, while automated prompts are generated using various search algorithms or other automated methods. The advantage of manual prompts is that they are typically well-designed and can capture important aspects of the downstream task. However, the manual prompt design process can be time-consuming and costly. On the other hand, automated prompts are generated more quickly and can potentially capture more task-specific information than manually crafted prompts. However, the effectiveness of automated prompts heavily depends on the quality of the search algorithms and the representation power of the language model.

In summary, both manual prompts and automated prompts have their advantages and disadvantages. Manual prompts provide greater control over the output, while automated prompts are more efficient and adaptable. Ultimately, the choice of the prompt will depend on the specific task and available resources.

4   Applications of Prompts

Prompt engineering enables language models to achieve a wide range of AI tasks in the medical field with high performance. With customized medical prompts, models can accomplish complex healthcare problems that are previously difficult without huge medical datasets and resources. In the medical field, the applications of prompt engineering include:

- Classification [19, 32, 43, 50, 53–57]: Medical prompts elicit predicted diagnoses, categories or tags from models. For example, "The patient's

CT scan results show: [tumor type]" can steer a model to generate "thymoma", etc. Classification prompts provide context for model predictions in the medical domain. They are useful for condition classification, diagnosis classification, etc.

- Generation [18, 38, 58–63]: Open-ended medical prompts give models freedom to generate creative samples like case reports, draft research papers, and so on, from their knowledge. For instance, "Here is a possible case report:" can lead to an original case report. Generation prompts allow unconstrained production of new medical data samples. They are applicable for creative medical tasks like clinical case modeling and medical article writing.

- Detection [14, 64, 65]: Medical prompts frame object or anomaly detection contexts and ask models to report entities detected. For example, "Detect any abnormal findings you observe in the attached CT images". Detection prompts imply what medical models need to identify. They are relevant for locating and classifying diseases, anatomical anomalies or other irregularities in medical images, videos, and documents.

- Augmentation [15, 41, 66, 67]: Medical prompts provide additional context to augment data for training language models. For example, we can input samples of all classes into language models like ChatGPT and prompt the model to generate samples that preserves semantic consistency with existing labelled data. The generated data, together with original samples, are used to train classifier models. Augmentation prompts adds details to expand medical dataset. They are useful for boosting model performance on certain medical tasks by synthesizing more data context.

- Question Answering [40, 68–74]: Medical prompts pose questions for models to comprehend and accurately respond to based on medical knowledge. For example, "What year was radiocontrast invented?" prompts the model to search its medical knowledge base and answer "1923" [XX]. Question answering prompts assess if medical models understand questions and can generate correct responses by finding relevant medical information or facts.

- Inference [44, 75, 76]: Medical prompts present scenarios and have models explain their clinical reasoning or diagnostic process. For example, "Here are two medical variables: A = [some values], B = [some values]. Explain how A affects B and why." The model must logically relate A and B based on medical knowledge. Reasoning prompts require coherent medical explanations and inference generation. They aim to test clinical logic, causality reasoning, and argumentation abilities of medical models.

In summary, prompt engineering provides a framework of adapting and improving language models for diverse medical applications. Prompt design is key to unlocking model capabilities for different medical problems.

4.1 Classification Task

Prompt engineering techniques show great promise for medical classification tasks when annotated data is scarce. Some studies apply prompt-based learning with pre-trained language models for clinical and mental health classification.

Medical image analysis   GPT4MIA utilizes the Generative Pre-trained Transformer (GPT) as a plug-and-play inference model for medical image analysis (MIA) classification tasks [53]. The authors theoretically justify using GPT-3, a large pre-trained language model, for MIA. They develop prompt engineering techniques like improved prompt structure design, sample selection, and prompt ordering to enhance GPT4MIA's efficiency and effectiveness in detecting prediction errors and improving accuracy for image classification. GPT4MIA, working with established vision models, shows strong performance on these tasks.

Clinical text classification   HealthPrompt proposes a novel prompt-based clinical NLP framework that applies prompt engineering to pre-trained language models for classifying clinical texts without training data [32]. Deep learning models need large annotated datasets, but these are lacking for clinical NLP. HealthPrompt addresses this by using prompt-based learning to tune pre-trained language models for new tasks by defining prompt templates instead of fine-tuning the models. An analysis of HealthPrompt using six pre-trained language models in a no-data setting shows that prompts effectively capture the context in clinical texts and achieve good performance without training data.

Mental health classification   ChatGPT, an LLM-based model, demonstrates strong zero-shot performance in classifying social media posts for three mental health tasks: stress detection, depression detection, and suicidality detection [19]. LLMs show promise for NLP mental health applications but need data, which ChatGPT addresses using prompt-based learning. ChatGPT outperforms a baseline model, indicating the potential of language models for mental health classification, especially when data is limited.

4.2 Generation Task

Medical image generation   Chambon et al. demonstrate how selective fine-tuning and meaningful evaluation enabled the Stable Diffusion generative model to generate medical images from clinical prompts [58]. By customizing Stable Diffusion for the medical domain through focusing fine-tuning and assessing performance with clinically tailored metrics, the authors translate domain knowledge into model capabilities for sensitive generation where data is scarce. Their methods and results highlight the promise of prompt engineering to impart specialized expertise to models for nuanced generation addressing real-world challenges. However, fully realizing this potential will require continued progress in techniques for model adaptation and domain-specific evaluation.

Medical text generation NapSS presents a "summarize-then-simplify" strategy to coherently simplify medical text [59]. NapSS generates summaries and narrative prompts to respectively train a model to extract key content and guide a generator to clarify it while preserving flow. Evaluated on medical text, NapSS improves on baselines in lexical, semantic, and human metrics. This highlights how multi-stage prompt engineering achieves nuanced, domain-relevant generation requiring sophisticated evaluation.

Medical report translation Lyu et al. examine the LLM ChatGPT's performance in translating radiology reports into plain language for education [18]. Evaluated by radiologists, ChatGPT shows promise but some inconsistency improvements with better prompts. ChatGPT generats both general and specific recommendations based on reports. Compared to GPT-4, which is a newer model, ChatGPT's performance is significantly outperformed, indicating LLMs' potential for clinical use but requiring advanced models and prompts for optimized performance.

4.3   Detection Task

Medical image detection Qin et. al show how prompt engineering enables vision-language models (VLM) pretrained on natural images to transfer knowledge to medical images [64]. Manual prompts containing expert medical knowledge improve VLM performance on medical detection tasks with no medical training data. Automatic prompts inject image details, further enhancing performance. Proposed approaches outperform default prompts on 13 medical datasets. Fine-tuned models beat supervised baselines, demonstrating VLMs can learn from limited medical data by transferring knowledge from natural images. This work establishes prompt engineering as key to applying VLMs in medicine and paves the way for developing VLMs tailored to healthcare applications.

Medical text detection DeID-GPT is the first framework to de-identify free text medical data using GPT-4's state-of-the-art NLP capabilities, which achieves the highest accuracy in removing private information while preserving meaning [14]. It requires no changes for different data types thanks to GPT-4's scale and in-context learning. DeID-GPT uses prompts to generate optimal de-identification results with little human input, showing the potential of ChatGPT and GPT-4 for automated medical text processing. It contributes a novel, highly effective approach to an important problem - enabling the use of medical text data while protecting patient privacy. DeID-GPT establishes GPT-4 and similar LLMs as a means to overcome grand challenges in healthcare through prompt engineering and natural language understanding.

4.4 Augmentation Task

Text data augmentation Dai et. al propose a new text data augmentation method called ChatAug [15]. ChatAug uses the ChatGPT language model to rephrase sentences in the training data into multiple conceptually similar but semantically different samples, increasing the sample size. They apply ChatAug to few-shot learning text classification in the medical domain and show improved performance over state-of-the-art data augmentation methods in terms of both testing accuracy and distribution of the generated samples.

Transform clinical data into prompts Milecki et. al propose a new multimodal learning model called MEDIMP - Medical Images and Prompts [66]. MEDIMP translates clinical biomedical data into text prompts which are input into LLMs to produce augmented data. Contrastive learning and image-text pairs are then used to learn meaningful representations of medical images, in this case renal transplant DCE MRI images. MEDIMP generates prompts using predefined sentence templates and ChatGPT, aiming to learn representations useful for prognosis of patient status 2-4 years after transplant.

4.5 Question Answering Task

Medical text-based QA Singhal et. al present medical QA datasets, Multi-MedQA and HealthSearchQA, and proposes a framework for evaluating clinical LLMs along dimensions like factuality, precision, harm and bias [68]. To adapt LLMs like PaLM and FlanPaLM for the medical domain, the authors apply prompt engineering techniques like instruction prompt tuning. A parameter-efficient prompting approach for aligning LLMs to specialized domains using a few examples. The resulting model, Med-PaLM, shows encouraging improvements but still lags clinicians. This work highlights the importance of medical QA datasets and human-centered evaluation in creating beneficial clinical language models. By systematically evaluating tuned language models, the authors reveal gaps in comprehending medical knowledge and reasoning that must be addressed to develop models for healthcare.

Medical image-based QA The open-ended medical visual QA work treats the task as generative. Sonsbeek et. al develop a network to map visual features to tokens that, with the question, directly prompt a PLM [69]. Exploring PLM fine-tuning strategies, their approach generates open-ended responses that outperform others on medical QA benchmarks like Slake, OVQA and PathVQA. This enables a PLM to understand medical images for open-ended medical visual QA where answers are not constrained to a predefined set. The work points to promising directions for open-domain medical visual question answering using prompt-based generative models.

Medical video-based QA Visual-Prompt Text Span Localization (VPTSL) proposes a novel approach to the temporal answering grounding in video (TAGV)

task by formulating it as predicting the span of timestamped subtitles that matches the visual answer [70]. To bridge the semantic gap between textual questions and visual answers, VPTSL introduces visual prompts—highlight features obtained from video-text highlighting using the question. These visual prompts are fed into the PLM along with the subtitles and question. A text span predictor then models these visual and textual prompts to predict the subtitle span. VPTSL outperforms the state-of-the-art on MedVidQA by a large margin (28.36% in mIOU), showing the effectiveness of visual prompts and the text span predictor for medical video QA. This work enables PLMs to ground temporal answers in instructional videos by prompting them with contextual visual information.

4.6   Inference Task

Natural language inference in radiology   Wu et. al evaluate ChatGPT and GPT-4 on a radiology natural language inference task [75]. They implement zero-shot and few-shot prompts for the models. The zero-shot prompt provides only task instructions and sentence-question pairs, requiring the models to determine entailment with no labeled examples. The few-shot prompt incorporates 10 labeled sentence-question examples to provide in-context learning before the evaluation pair. By designing these prompts with varying levels of contextual information, the authors show that ChatGPT and GPT-4 can achieve over 50% accuracy on the radiology task without requiring large amounts of training data. GPT-4 outperforms ChatGPT, indicating its greater capability. The results demonstrate the feasibility of building generic models that can perform well across different domains.

Causal reasoning about medical variables   Long et. al examine whether GPT-3 can accurately predict the presence or absence of edges between variables in causal graphs based on medical context [76]. They evaluate GPT-3 using different prompts (e.g. declarative v.s. interrogative) and linking verbs (e.g. "causes" v.s. "correlates with"). They find GPT-3's performance varies based on these factors, indicating its sensitivity to user input. With well-designed prompts using "causes", GPT-3 achieves over 50% accuracy on the tested graphs. Further, the authors select 3 causal graphs of different complexities from the medical literature. For each graph, they generate 2000 randomly permuted sentence pairs from the graph. GPT-3 predicts whether a causal edge exists between each pair. Its accuracy is highest (70%-85%) on the simplest graph. Though performance decreases for the complex graphs, it remains well above the 50% random baseline, demonstrating GPT-3's potential for reasoning about medical variables.

4.7 A Recap of Prompt Engineering for Classification, Generation, Detection, Augmentation, QA, and Inference

These studies highlight the potential of using prompt engineering for pre-trained language models to advance clinical NLP and a range of medical AI applications, especially when annotated data is scarce. Prompt-based learning tunes models for new tasks by defining task templates instead of fine-tuning, helping models achieve strong performance in zero-shot learning scenarios where they generalize to new classes without examples. Large pre-trained language models provide an ideal initialization for prompt engineering across medical domains and tasks when combined with effective prompt design.

Though more work is needed, these papers demonstrate the promise of prompt engineering and LLMs for medical AI with limited data. Prompt-based learning could help address lacks of annotated data and advance applications like classification, generation, detection, enhancement, question answering, and reasoning by unlocking models' capabilities through tailored prompts. Strong zero-shot performance suggests prompt engineering may reduce data needs for medical AI versus traditional supervised learning.

5 Challenges and Future Directions

This chapter discusses the challenges, current research directions, and future opportunities and development directions of prompt-based methods.

Firstly, the challenges in prompt engineering are addressed, including the data scarcity [77] in the medical NLP domain, the interpretability of models, and inherent issues in prompt engineering.

Secondly, the current research directions are introduced, which include prompt generation [1, 78], prompt optimization [49], multimodal data processing [60], and deep reinforcement learning [79]. These research directions aim to improve the effectiveness and applicability of prompt-based methods, further promoting the development of the NLP field.

Lastly, the opportunities and future development directions of Prompt-based methods are explored, such as the development of multitask and multimodal processing and the expansion of innovative Prompt design and intelligent Prompt generation [2, 49]. These opportunities and development directions provide a broad space and prospects for the future development of Prompt-based methods.

5.1 Challenges in Prompt Engineering

When it comes to NLP tasks in the medical field, prompt engineering faces several challenges. First, medical data is usually limited and specialized, making it difficult to cover all domain knowledge and thus restricting the model's generalization ability [14, 60]. Secondly, the medical field is rich in terminology and domain knowledge, but often complex, which requires effective integration

of this knowledge into prompts [12]. Additionally, different medical tasks require different prompt designs, which requires balancing the complexity and interpretability of prompts while also utilizing existing medical domain knowledge to guide the model to generate high-quality predictions [68]. Finally, the prompt design also needs to consider how to avoid introducing any human bias or erroneous information to ensure the fairness and accuracy of the model [80].

It should be noted that NLP tasks in the medical field are highly challenging as they involve a large amount of domain knowledge and specialized terminology, which may not be easily understood or processed by general NLP models [81]. Therefore, the importance of prompt engineering in the medical field is self-evident. These challenges include, but are not limited to:

Data scarcity: In the medical field, many tasks require the use of specific medical data, which is often very scarce and difficult to obtain [14, 82]. Moreover, the special nature of medical data involves ethical issues, which poses a great challenge to prompt engineering. The lack of sufficient data makes it difficult to design accurate and effective prompts.

Data uncertainty: The medical field involves numerous domain knowledge and terminology, which may have different interpretations and usage in different texts, leading to increased uncertainty in prompt design [73, 82].

Model interpretability: In the medical field, model interpretability is particularly important. As it involves human health and life, the model's prediction results need to be reasonably explained and justified. Therefore, in the medical field, prompt design not only considers the accuracy of the model but also its interpretability [14].

Additionally, prompt engineering in the medical field faces common challenges in other fields, such as prompt engineering's self-consistency issues [83], prompt leakage [84], and adversarial issues [85], among others. These challenges in medical tasks are more severe than in other fields. Future research needs to fully consider these challenges and propose corresponding solutions.

In summary, prompt engineering in the medical field is critical to the success of NLP tasks, but it also faces various challenges, such as data scarcity[86, 87] and uncertainty[88], model interpretability[89, 90], self-consistency[91], and adversarial issues[92]. Addressing these challenges requires innovative solutions and a comprehensive understanding of the medical domain knowledge and its related terminology.

5.2   Current Research Directions

As the development and application of artificial general intelligence (AGI) models continue to advance, researchers in the medical field have also begun to apply them to various medical tasks. Currently, the research direction of prompt engineering in the medical field is very diverse, covering different types of prompt designs [60], multi-modal data processing [57], and deep reinforcement learning [93], among other aspects.

In terms of prompt design methods, it is necessary to consider the specific characteristics of the task and data features to make appropriate choices. These

include simple manually designed template-style prompts, prompts based on knowledge graphs[94], prompts generated based on natural language[95], and prompts that can be embedded as trainable parameters in the model for automatic generation[96]. In multi-modal data processing, integrating text and image data requires the design of appropriate prompts to guide the model to handle and analyze different types of data, thereby improving the model's performance and interpretability[97]. Additionally, deep reinforcement[98] learning can continuously optimize prompt design through autonomous learning, further improving the performance of the model.

In summary, the current research direction of prompt engineering in the medical field is diverse, and researchers can choose appropriate methods and techniques based on the requirements of the task and data characteristics to improve the performance and interpretability of the model.

5.3   Opportunities and Future Directions

In the field of medicine, prompt engineering presents a wide range of opportunities and future directions for application. The development of LLMs such as ChatGPT and GPT-4 provides tremendous potential for prompt engineering in the medical field [99]. As an essential part of prompt-based technology, prompt engineering can assist large models in understanding and processing medical data more accurately [100], from manual prompts to automated prompts, improving the automation and intelligence of prompts [101]. Applying automated algorithms to prompt engineering, utilizing data-driven methods to generate more intelligent and efficient prompts, can enhance the efficiency and accuracy of NLP tasks in the medical field [102]. With the emergence of multimodal information in the medical field, combining multiple modalities such as text, images, and speech can better address practical problems in the medical domain[103]. Additionally, conducting multi-task research by integrating multiple medical research tasks can provide better services for clinical healthcare[104, 105]. In conclusion, prompt engineering is a promising area for medical NLP research, and future studies should focus on developing more intelligent and automated prompt engineering methods and exploring the combination of multiple modalities and tasks.

6   Conclusion

In this section, we will provide a summary of the review and discuss its limitations and contributions.

6.1   Summary of the Review

Overall, this review article provides a comprehensive overview of the different prompt engineering methods and challenges in the context of NLP tasks in the medical field. We have presented different prompt design methods, including

manual and automated, discrete and continuous, and provided examples to illustrate their practical applications in medical settings. Additionally, we have discussed the different prompt engineering methods used for various medical tasks, such as classification, generation, detection, argumentation, reconstruction, question answering, prediction, and inference tasks.

One key finding of this review is that prompt engineering is a promising approach to improving the performance of NLP tasks in the medical field. By carefully designing prompts that are tailored to specific tasks and domains, researchers can achieve significant improvements in accuracy and efficiency.

However, there are also several challenges and limitations to prompt engineering that need to be addressed in future research. Moreover, the effectiveness of prompts may depend on the specific characteristics of the task and the domain, which can vary widely across different medical applications.

6.2   Limitations of the Review

Despite the comprehensive literature search process and the inclusion of 333 relevant studies, it is possible that some relevant research in the field of medical NLP prompt engineering may have been missed. Therefore, the scope of this review may not be entirely exhaustive, and there could be other important studies that are not covered.

Additionally, although the article covers a range of medical tasks and the prompt engineering methods employed in each of them, there may be other tasks or research methodologies that are not explored. Future research in this field should, therefore, continue to investigate the potential of prompt engineering in other areas of medical NLP and explore additional methods that may not have been considered in this review.

6.3   Conclusion and Contributions

This review article makes a significant contribution to the field by providing a comprehensive and systematic overview of prompt engineering methods for NLP tasks in the medical domain. By covering a wide range of methods from manual to automated, and from discrete to continuous, the article serves as a valuable reference and guide for researchers in the medical NLP field to help them choose and design prompts that can improve model performance and efficiency. Additionally, the article discusses the unique challenges posed by medical NLP tasks and explores the future research directions for prompt engineering in this field. Overall, this review article is a valuable resource for medical NLP researchers and provides insightful guidance for future research in this rapidly growing field.

6.4   Recommendations for Future Research

In summary, the future of NLP in the medical field looks promising, with more advanced techniques such as prompt engineering and multi-modal integration being explored. We can expect to see more research in applying state-of-the-art

models like ChatGPT and GPT-4 to medical NLP tasks, as well as developing more efficient and accurate prompt engineering methods[106]. Furthermore, there is a need to explore other areas of NLP in medical[75], such as text-based medical image analysis and building medical knowledge graphs.

As the field continues to evolve, the development of large models such as GPT-4 will further advance NLP research. However, this will also require innovation and breakthroughs in both hardware and software to accommodate the higher computing and resource demands.

Overall, future work in this field will require a comprehensive and precise approach that incorporates multi-modal information and large model technology to advance medical NLP research and applications. With the potential for cross-disciplinary collaboration between medical professionals and NLP researchers, the possibilities for improving healthcare through NLP are vast.

References

[1] Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys. 2023;55(9):1-35.

[2] White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:230211382. 2023.

[3] Hu M, Pan S, Li Y, Yang X. Advancing Medical Imaging with Language Models: A Journey from N-grams to ChatGPT. arXiv preprint arXiv:230404920. 2023.

[4] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research. 2020;21(1):5485-551.

[5] Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: A survey. Science China Technological Sciences. 2020;63(10):1872-97.

[6] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:210408691. 2021.

[7] Kopač J, Bahor M, Soković M. Optimal machining parameters for achieving the desired surface roughness in fine turning of cold pre-formed steel workpieces. International Journal of Machine Tools and Manufacture. 2002;42(6):707-16.

[8] Ding N, Hu S, Zhao W, Chen Y, Liu Z, Zheng H, et al. OpenPrompt: An Open-source Framework for Prompt-learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2022. p. 105-13.

[9] Liu V, Chilton LB. Design guidelines for prompt engineering text-to-image generative models. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; 2022. p. 1-23.

[10] Sorensen T, Robinson J, Rytting C, Shaw A, Rogers K, Delorey A, et al. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022. p. 819-62.

[11] Rong G, Mendez A, Assi EB, Zhao B, Sawan M. Artificial intelligence in healthcare: review and prediction case studies. Engineering. 2020;6(3):291-301.

[12] Zhou L, Hripcsak G. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. Journal of biomedical informatics. 2007;40(2):183-202.

[13] Baud R, Rassinoux AM, Scherrer JR. Natural language processing and semantical representation of medical texts. Methods of information in medicine. 1992;31(02):117-25.

[14] Liu Z, Yu X, Zhang L, Wu Z, Cao C, Dai H, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:230311032. 2023.

[15] Dai H, Liu Z, Liao W, Huang X, Wu Z, Zhao L, et al. Chataug: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:230213007. 2023.

[16] Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y, et al. Language Models as Knowledge Bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. p. 2463-73.

[17] Wei X, Cui X, Cheng N, Wang X, Zhang X, Huang S, et al. Zero-Shot Information Extraction via Chatting with ChatGPT. arXiv preprint arXiv:230210205. 2023.

[18] Lyu Q, Tan J, Zapadka ME, Ponnatapuram J, Niu C, Wang G, et al. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: Promising results, limitations, and potential. arXiv preprint arXiv:230309038. 2023.

[19] Lamichhane B. Evaluation of ChatGPT for NLP-based Mental Health Applications. arXiv preprint arXiv:230315727. 2023.

[20] Wallace E, Feng S, Kandpal N, Gardner M, Singh S. Universal Adversarial Triggers for Attacking and Analyzing NLP. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. .

[21] Shin T, Razeghi Y, Logan IV RL, Wallace E, Singh S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020. p. 4222-35.

[22] Li XL, Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 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); 2021. p. 4582-97.

[23] Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. International Journal of Computer Vision. 2022;130(9):2337-48.

[24] Kenton JDMWC, Toutanova LK. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT; 2019. p. 4171-86.

[25] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 16000-9.

[26] Liu PJ, Saleh M, Pot E, Goodrich B, Sepassi R, Kaiser L, et al. Generating Wikipedia by Summarizing Long Sequences. In: International Conference on Learning Representations;. .

[27] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019;1(8):9.

[28] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[29] Najar A, Chetouani M. Reinforcement learning with human advice: a survey. Frontiers in Robotics and AI. 2021;8:584075.

[30] Knox WB, Stone P. Interactively shaping agents via human reinforcement: The TAMER framework. In: Proceedings of the fifth international conference on Knowledge capture; 2009. p. 9-16.

[31] Cui L, Wu Y, Liu J, Yang S, Zhang Y. Template-Based Named Entity Recognition Using BART. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021; 2021. p. 1835-45.

[32] Sivarajkumar S, Wang Y. HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing. arXiv preprint arXiv:220305061. 2022.

[33] Milecki L, Kalogeiton V, Bodard S, Anglicheau D, Correas JM, Timsit MO, et al. MEDIMP: Medical Images and Prompts for renal transplant representation learning. In: Medical Imaging with Deep Learning;. .

[34] Feng W, Bu X, Zhang C, Li X. Beyond Bounding Box: Multimodal Knowledge Learning for Object Detection. arXiv preprint arXiv:220504072. 2022.

[35] Yang C, Woicik A, Poon H, Wang S. BLIAM: Literature-based Data Synthesis for Synergistic Drug Combination Prediction. arXiv preprint arXiv:230206860. 2023.

[36] Liang S, Zhao M, Schütze H. Modular and Parameter-Efficient Multimodal Fusion with Prompting. arXiv preprint arXiv:220308055. 2022.

[37] Gao Z, Hu Y, Tan C, Li SZ. PrefixMol: Target-and Chemistry-aware Molecule Design via Prefix Embedding. arXiv preprint arXiv:230207120. 2023.

[38] Abaho M, Bollegala D, Williamson P, Dodd S. Position-based Prompting for Health Outcome Generation. In: Proc. of The 21st BioNLP workshop associated with the ACL SIGBIOMED special interest group; 2022. .

[39] Lai VD, Ngo NT, Veyseh APB, Man H, Dernoncourt F, Bui T, et al. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. arXiv preprint arXiv:230405613. 2023.

[40] Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. arXiv preprint arXiv:230401938. 2023.

[41] Yuan J, Tang R, Jiang X, Hu X. LLM for Patient-Trial Matching: Privacy-Aware Data Augmentation Towards Better Performance and Generalizability. arXiv preprint arXiv:230316756. 2023.

[42] Zhou Y, Muresanu AI, Han Z, Paster K, Pitis S, Chan H, et al. Large language models are human-level prompt engineers. arXiv preprint arXiv:221101910. 2022.

[43] Wang Y, Deng J, Wang T, Zheng B, Hu S, Liu X, et al. Exploiting prompt learning with pre-trained language models for Alzheimer's Disease detection. arXiv preprint arXiv:221016539. 2022.

[44] Wang H, Liu C, Xi N, Zhao S, Ju M, Zhang S, et al. Prompt Combines Paraphrase: Teaching Pre-trained Models to Understand Rare Biomedical Words. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022. p. 1422-31.

[45] Romanov A, Shivade C. Lessons from Natural Language Inference in the Clinical Domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018. p. 1586-96.

[46] Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. In: Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021. Association for Computational Linguistics (ACL); 2021. p. 3816-30.

[47] Chen T, Liu L, Jia X, Cui B, Tang H, Tang S. Distilling Task-specific Logical Rules from Large Pre-trained Models. arXiv preprint arXiv:221002768. 2022.

[48] Ruan Y, Lan X, Tan DJ, Abdullah HR, Feng M. Medical Intervention Duration Estimation Using Language-enhanced Transformer Encoder with Medical Prompts. arXiv preprint arXiv:230317408. 2023.

[49] Ma C, Wu Z, Wang J, Xu S, Wei Y, Liu Z, et al. ImpressionGPT: An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT. arXiv preprint arXiv:230408448. 2023.

[50] Elfrink A, Vagliano I, Abu-Hanna A, Calixto I. Soft-prompt tuning to predict lung cancer using primary care free-text Dutch medical notes. arXiv preprint arXiv:230315846. 2023.

[51] Wang Y, Huang W, Li C, Zheng X, Lin Y, Wang S. Adaptive PromptNet For Auxiliary Glioma Diagnosis without Contrast-Enhanced MRI. arXiv preprint arXiv:221107966. 2022.

[52] Li Z, Zhao L, Zhang Z, Zhang H, Liu D, Liu T, et al. Steering Prototype with Prompt-tuning for Rehearsal-free Continual Learning. arXiv preprint arXiv:230309447. 2023.

[53] Zhang Y, Chen DZ. GPT4MIA: Utilizing Geneative Pre-trained Transformer (GPT-3) as A Plug-and-Play Transductive Model for Medical Image Analysis. arXiv preprint arXiv:230208722. 2023.

[54] Akrout M, Gyepesi B, Holló P, Poór A, Kincső B, Solis S, et al. Diffusion-based Data Augmentation for Skin Disease Classification: Impact Across Original Medical Datasets to Fully Synthetic Images. arXiv preprint arXiv:230104802. 2023.

[55] Zhu W, Zhou R, Yuan Y, Timothy C, Jain R, Luo J. SegPrompt: Using Segmentation Map as a Better Prompt to Finetune Deep Models for Kidney Stone Classification. arXiv preprint arXiv:230308303. 2023.

[56] Chen Z, Diao S, Wang B, Li G, Wan X. Towards Unifying Medical Vision-and-Language Pre-training via Soft Prompts. arXiv preprint arXiv:230208958. 2023.

[57] Lin Y, Zhao Z, ZHU Z, Wang L, Cheng KT, Chen H. Exploring Visual Prompts for Whole Slide Image Classification with Multiple Instance Learning. arXiv preprint arXiv:230313122. 2023.

[58] Chambon P, Bluethgen C, Langlotz CP, Chaudhari A. Adapting pre-trained vision-language foundational models to medical imaging domains. arXiv preprint arXiv:221004133. 2022.

[59] Lu J, Li J, Wallace BC, He Y, Pergola G. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. arXiv preprint arXiv:230205574. 2023.

[60] Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:230207257. 2023.

[61] Lehman E, Lialin V, Legaspi KY, Sy AJR, Pile PTS, Alberto NRI, et al. Learning to Ask Like a Physician. arXiv preprint arXiv:220602696. 2022.

[62] Weber T, Ingrisch M, Bischl B, Rügamer D. Cascaded Latent Diffusion Models for High-Resolution Chest X-ray Synthesis. arXiv preprint arXiv:230311224. 2023.

[63] Lee S, Lee DY, Im S, Kim NH, Park SM. Clinical Decision Transformer: Intended Treatment Recommendation through Goal Prompting. arXiv preprint arXiv:230200612. 2023.

[64] Qin Z, Yi H, Lao Q, Li K. Medical Image Understanding with Pre-trained Vision Language Models: A Comprehensive Study. arXiv preprint arXiv:220915517. 2022.

[65] Ye Y, Xie Y, Zhang J, Chen Z, Xia Y. UniSeg: A Prompt-driven Universal Segmentation Model as well as A Strong Representation Learner. arXiv preprint arXiv:230403493. 2023.

[66] Milecki L, Kalogeiton V, Bodard S, Anglicheau D, Correas JM, Timsit MO, et al. MEDIMP: Medical Images and Prompts for renal transplant representation learning. arXiv preprint arXiv:230312445. 2023.

[67] Li B, Wang X, Xu X, Hou Y, Feng Y, Wang F, et al. Semantic-Guided Image Augmentation with Pre-trained Models. arXiv preprint arXiv:230202070. 2023.

[68] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge. arXiv preprint arXiv:221213138. 2022.

[69] van Sonsbeek T, Derakhshani MM, Najdenkoska I, Snoek CG, Worring M. Open-Ended Medical Visual Question Answering Through Prefix Tuning of Language Models. arXiv preprint arXiv:230305977. 2023.

[70] Li B, Weng Y, Sun B, Li S. Towards visual-prompt temporal answering grounding in medical instructional video. arXiv preprint arXiv:220306667. 2022.

[71] Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. arXiv preprint arXiv:230318027. 2023.

[72] Jang D, Kim CE. Exploring the Potential of Large Language models in Traditional Korean Medicine: A Foundation Model Approach to Culturally-Adapted Healthcare. arXiv preprint arXiv:230317807. 2023.

[73] Zuccon G, Koopman B. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. arXiv preprint arXiv:230213793. 2023.

[74] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:230313375. 2023.

[75] Wu Z, Zhang L, Cao C, Yu X, Dai H, Ma C, et al. Exploring the Trade-Offs: Unified Large Language Models vs Local Fine-Tuned Models for Highly-Specific Radiology NLI Task. arXiv preprint arXiv:230409138. 2023.

[76] Long S, Schuster T, Piché A, Research S, et al. Can large language models build causal graphs? arXiv preprint arXiv:230305279. 2023.

[77] Azaria A, Ekblaw A, Vieira T, Lippman A. Medrec: Using blockchain for medical data access and permission management. In: 2016 2nd international conference on open and big data (OBD). IEEE; 2016. p. 25-30.

[78] He Y, Zheng S, Tay Y, Gupta J, Du Y, Aribandi V, et al. Hyperprompt: Prompt-based task-conditioning of transformers. In: International Conference on Machine Learning. PMLR; 2022. p. 8678-90.

[79] François-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J, et al. An introduction to deep reinforcement learning. Foundations and Trends® in Machine Learning. 2018;11(3-4):219-354.

[80] Crothers E, Japkowicz N, Viktor H. Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. arXiv preprint arXiv:221007321. 2022.

[81] Chen Z, Zhou Q, Shen Y, Hong Y, Zhang H, Gan C. See, Think, Confirm: Interactive Prompting Between Vision and Language Models for Knowledge-based Visual Reasoning. arXiv preprint arXiv:230105226. 2023.

[82] Yao Z, Cao Y, Yang Z, Deshpande V, Yu H. Extracting Biomedical Factual Knowledge Using Pretrained Language Model and Electronic Health Record Context. arXiv preprint arXiv:220907859. 2022.

[83] Wang P, Gao H, Guo X, Xiao C, Qi F, Yan Z. An Experimental Investigation of Text-based CAPTCHA Attacks and Their Robustness. ACM Computing Surveys. 2023;55(9):1-38.

[84] Perez F, Ribeiro I. Ignore Previous Prompt: Attack Techniques For Language Models. In: NeurIPS ML Safety Workshop;. .

[85] McGuffie K, Newhouse A. The radicalization risks of GPT-3 and advanced neural language models. arXiv preprint arXiv:200906807. 2020.

[86] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nature medicine. 2022;28(1):31-8.

[87] Liu X, Zhao J, Li J, Cao B, Lv Z. Federated neural architecture search for medical data security. IEEE transactions on industrial informatics. 2022;18(8):5628-36.

[88] Wang K, Zhan B, Zu C, Wu X, Zhou J, Zhou L, et al. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Medical Image Analysis. 2022;79:102447.

[89] Barragan-Montero AM, Bibal A, Huet M, Draguet C, Valdes G, Nguyen D, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. Physics in Medicine & Biology. 2022.

[90] Li X, Xiong H, Li X, Wu X, Zhang X, Liu J, et al. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems. 2022;64(12):3197-234.

[91] Wang X, Wei J, Schuurmans D, Le Q, Chi E, Zhou D. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:220311171. 2022.

[92] Rasool RU, Ahmad HF, Rafique W, Qayyum A, Qadir J. Security and privacy of internet of medical things: A contemporary review in the age of surveillance, botnets, and adversarial ML. Journal of Network and Computer Applications. 2022:103332.

[93] Zhong T, Wei Y, Yang L, Wu Z, Liu Z, Wei X, et al. ChatABL: Abductive Learning via Natural Language Interaction with ChatGPT. arXiv preprint arXiv:230411107. 2023.

[94] Andrus BR, Nasiri Y, Cui S, Cullen B, Fulda N. Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36; 2022. p. 10436-44.

[95] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does CHATGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment. JMIR Medical Education. 2023;9(1):e45312.

[96] Niu C, Wang G. CT Multi-Task Learning with a Large Image-Text (LIT) Model. bioRxiv. 2023:2023-04.

[97] Behrad F, Abadeh MS. An overview of deep learning methods for multimodal medical data mining. Expert Systems with Applications. 2022:117006.

[98] Jacob B, Kaushik A, Velavan P, Sharma M. Autonomous drones for medical assistance using reinforcement learning. Advances in Augmented Reality and Virtual Reality. 2022:133-56.

[99] Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:230401852. 2023.

[100] Chen JS, Baxter SL. Applications of natural language processing in ophthalmology: present and future. Frontiers in Medicine. 2022;9.

[101] Thakur M, Dhanalakshmi S, Kuresan H, Senthil R, Narayanamoorthi R, Lai KW. Automated restricted Boltzmann machine classifier for early diagnosis of Parkinson's disease using digitized spiral drawings. Journal of Ambient Intelligence and Humanized Computing. 2023;14(1):175-89.

[102] Martin EA, D'Souza AG, Lee S, Doktorchik C, Eastwood CA, Quan H. Hypertension identification using inpatient clinical notes from electronic medical records: an explainable, data-driven algorithm study. Canadian Medical Association Open Access Journal. 2023;11(1):E131-9.

[103] Azam MA, Khan KB, Salahuddin S, Rehman E, Khan SA, Khan MA, et al. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Computers in biology and medicine. 2022;144:105253.

[104] Chengoden R, Victor N, Huynh-The T, Yenduri G, Jhaveri RH, Alazab M, et al. Metaverse for Healthcare: A Survey on Potential Applications, Challenges and Future Directions. IEEE Access. 2023.

[105] Buchlak QD, Esmaili N, Bennett C, Farrokhi F. Natural language processing applications in the clinical neurosciences: A machine learning augmented systematic review. Machine Learning in Clinical Neuroscience: Foundations and Applications. 2022:277-89.

[106] Kutela B, Msechu K, Das S, Kidando E. Chatgpt's Scientific Writings: A Case Study on Traffic Safety. Available at SSRN 4329120. 2023.