MARS: MedicAl thRead Summarization Dataset based on IIYI with Comparative Analysis of Large Language Models

Ruiru Zhang

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2024

Reading Committee:
Fei Xia, Chair
Wen-wai Yim

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

MARS: MedicAl thRead Summarization Dataset based on IIYI with Comparative Analysis of Large Language Models

Ruiru Zhang

Chair of the Supervisory Committee:
Professor Fei Xia
Department of Linguistics

This thesis presents MARS (**M**edic**A**l th**R**ead **S**ummarization Dataset based on IIYI), a pioneering dataset designed for medical domain thread summarization. MARS features a structure that captures the complexities and nuances of medical dialogues. The dataset integrates information extraction and summarization tasks, enabling a comprehensive evaluation of language models (LLMs) through extracting relevant information and generating coherent summaries. It also introduces unique challenges that necessitate advanced reasoning from LLMs, reflecting the complexities of healthcare discussions where misunderstandings can impact patient care. Furthermore, MARS serves as a critical benchmark for assessing LLM performance in a medical context, addressing a significant gap in existing literature. In addition to constructing the dataset, we tested the performance of various large language models on MARS, emphasizing the advantages of the GLM-4-Plus model when utilizing dynamic few-shot learning strategies. The experimental results further indicate that an extraction-then-summarization approach significantly enhances summarization performance compared to direct summarization methods. By providing diverse examples pertinent to real-world medical inquiries, MARS aims to promote robust research and the development of LLMs tailored to the intricacies of medical discourse, ultimately enhancing healthcare applications.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express sincere appreciation to University of Washington and the faculty, staff, and fellow students of the CLMS program for their support. I am especially grateful to Professor Fei Xia and Dr. Wen-wai Yim from Microsoft for providing me with the opportunity to work on this project and for their invaluable guidance throughout the completion of this thesis. Additionally, I would like to extend my heartfelt thanks to my partner, Jinyu Cao, for his unwavering support and encouragement. This thesis marks the conclusion of my journey but serves as the beginning of my work in this field. I look forward to contributing further to research in the areas of LLMs and AI in the future.

DEDICATION

to my days in Seattle, Washington

Chapter 1

INTRODUCTION

Online forum summarization plays a valuable role in distilling complex discussions into concise summaries that capture the essence of user interactions and viewpoints, enabling users to quickly grasp the salient points of a conversation. While several datasets have been developed for summarizing general online discussions, such as ConvoSumm [20] and AnswerSumm [21], the field remains notably underexplored, especially within the medical domain. The scarcity of dedicated datasets for medical forum summarization presents a significant gap in the literature, as existing datasets often fail to address the unique challenges posed by healthcare discussions, which require careful attention to accurate medical terminology, context, and the high stakes surrounding patient information.

In the medical context, the summarization task is particularly important due to the potential implications for patient care and clinical outcomes. Notably, previous research highlights the complexity of medical dialogues [7] [8] [30], where it is essential to accurately represent medical conditions, treatment history, and other pertinent information while avoiding omissions or inaccuracies that could adversely affect patient safety. Our work aims to address this critical gap by introducing a novel dataset specifically designed for summarizing medical forum discussions, contributing to the advancement of automated summarization techniques in a domain where clarity and precision are paramount. This dataset will not only facilitate further research in medical discourse but also aid in the development of applications that improve information accessibility in healthcare settings.

In this study, we present the development of the MARS (MedicAl thRead Summarization Dataset based on IIYI (爱爱医)) dataset, which is built upon the Chinese discussion data sourced from the IIYI medical forum. The IIYI platform serves as an important resource

for patients and medical professionals, enabling users to submit queries accompanied by relevant files and images. Our dataset specifically focuses on discussions about dermatology and sexually transmitted diseases(皮肤及性传播疾病讨论版), capturing a diverse array of interactions that reflect real-world medical inquiries. We manually conduct rigorous double annotations under a carefully designed guideline and automatically generate the summaries based on the extractive results.

The MARS dataset is structured from three aspects. First, it features the cleaned forum content shared with DermaVQA [67] (see Figure 1.1), which serves as the foundation for subsequent analyses. Second, we cautiously annotate the data and provide the annotation results from the perspective of information extraction that go beyond conventional entity and relation extraction methods (as shown in Table 1.1. This aspect involves identifying complex phrases related to treatment and diagnosis, presenting a more nuanced view of the medical discourse. Finally, the dataset includes structured summaries derived from the annotations provided by trained annotators (as shown in Figure 1.2). These summaries encapsulate the essential elements of each discussion, providing clear and organized insights into patient conditions, diagnoses, treatment plans, and recommendations. By thoughtfully crafting the MARS dataset, we aim to facilitate advances in information extraction and summarization within the medical domain, ultimately supporting the development of more effective healthcare applications and research methodologies.

In this study, we also aim to benchmark the performance of various large language models on MARS, focusing on their capabilities in information extraction and structured summarization. Our investigations are guided by two central research questions: First, we explore how effectively LLMs of varying sizes can generate structured summaries of medical thread discussions, considering the implications of zero-shot and dynamic few-shot learning strategies on their summarization performance. Second, we examine whether a sequential extraction followed by a summarization approach produces better results compared to direct summarization methods.

To assess these hypotheses, we evaluate five prominent LLMs, GPT-4o-mini [46], GLM-

Title: 是白癜风吗，请看图	**Title:** Is it Vitiligo? Please Check the Picture
Post: 病人中年妇女，五十岁左右，手背部出现暗红色皮疹，逐渐演变成白斑，局部摩擦后发红，病人无自觉症状，但有精神病表现，分别按白癜风及湿疹治疗，病情逐渐加重，皮疹变大，一月后面部出现淡红色斑疹，无鳞屑，面部无照片，拒绝病理	**Post:** The patient is a middle-aged woman, around fifty years old, with dark red rashes appearing on the back of her hands, gradually evolving into white patches. After local friction, the affected area turns red. The patient has no noticeable symptoms but exhibits psychiatric manifestations. Treatments for vitiligo and eczema were attempted, but the condition progressively worsened, with the rashes enlarging. A month later, light red rashes appeared on the face, with no scaling. There are no facial photographs available, and the patient refuses a biopsy.
Comments:	**Comments:**
*U00761:*暗红色皮疹，逐渐演变成白斑,考虑炎症后色素脱失。白癜风一出现就是白的,而且边界清楚	*U00761:* Dark red rashes gradually evolving into white patches suggest post-inflammatory hypopigmentation. Vitiligo appears white from the onset and has clear boundaries.
U00089: 是继发性白斑。多为暂时的，一般不需治疗，经半年至一年左右，可自行恢复。	*U00089:* It is secondary leukoderma. This is often temporary and generally does not require treatment. It can recover spontaneously in about six months to a year.
U01175: 白癜风一出现就是白的,而且边界清楚	*U01175:* Vitiligo appears white from the onset and has clear boundaries.
U00183: 患处是不是受过什么伤啊。的确像是炎症后色素脱失。	*U00183:* Was the affected area injured in some way? It indeed seems like post-inflammatory hypopigmentation.
U01101: 不是白癜风	*U01101:* It is not vitiligo.

Figure 1.1: An example of an original discussion thread sample from the MARS dataset: the left side shows the content of the original thread, while the right side presents its corresponding English translation. The original data is in Chinese. The "Uxxxxx" id refers to the user ID.

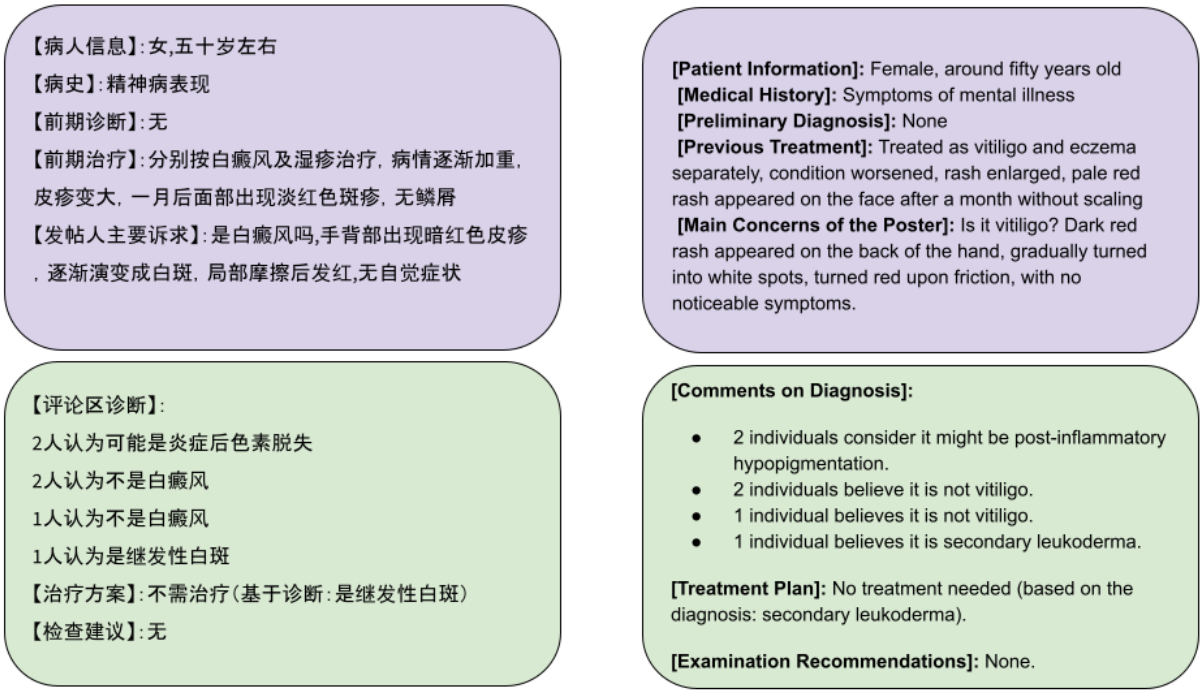

Figure 1.2: The automatically generated summary for the original data in Figure 1.1: the left side shows the summary of the original thread, while the right side presents its corresponding English translation. The purple block is the summary for the poster section, and the green block is the summary for the comment section. Summaries are generated based on the annotations shown in Table 1.1. Refer to the summary composition method in subsequent sections for more details.

turn_id	label	content	content_translated	Val	event
1	Problem_Current	是白癜风吗	Is it vitiligo?		E1
2	Sex	女	Female		
2	Age	五十岁左右	Around fifty years old		
2	Problem_Current	手背部出现暗红色皮疹，逐渐演变成白斑，局部摩擦后发红	Dark red rash appeared on the back of the hand, gradually turning into white spots, turning red upon friction		E1
2	Problem_Current	无自觉症状	No noticeable symptoms		E1
2	Clinical_History	精神病表现	Symptoms of mental illness		
2	Previous_Treatment	分别按白癜风及湿疹治疗，病情逐渐加重，皮疹变大， 一月后面部出现淡红色斑疹，无鳞屑	Treated as vitiligo and eczema separately, condition worsened, rash enlarged, pale red rash appeared on the face after a month without scaling		
3	Diagnosis	炎症后色素脱失	Post-inflammatory hypopigmentation	possible	E2
3	Diagnosis	白癜风	Vitiligo	hypothetical_negated	E3
4	Diagnosis	继发性白斑	Secondary leukoderma	present	E4
4	Treatment	不需治疗	No treatment needed		E4
5	Diagnosis	白癜风	Vitiligo	hypothetical_negated	E5
6	Diagnosis	炎症后色素脱失	Post-inflammatory hypopigmentation	possible	E6
7	Diagnosis	白癜风	Vitiligo	negated	E7

Table 1.1: The information extraction annotation for the original data in Figure 1.1: The column **content_translated** contains the English translations of **content** rather than annotations from English data. **turn_id** corresponds to the turn ID in a thread discussion, enabling linkage to the associated user ID for information tracking. The **Val** column indicates the certainty of the diagnosis as assessed by the commenter. The **event** column helps annotate relationships, indicating that annotations sharing the same event belong to a related context. Refer to the annotation guideline in subsequent sections for more details.

4-Plus [11], GLM-4-Flash [11], Qwen2-7B-Instruct [52], and Qwen2-72B-Instruct [52], considering their unique strengths in handling complex medical discussions that often feature extensive and diverse information. By implementing in-context learning strategies, including zero-shot and dynamic few-shot learning, we investigate how these methodologies impact the models' ability to accurately extract pertinent information and generate coherent summaries.

Our evaluation framework employs rigorous automatic assessment protocols, utilizing metrics such as ROUGE and BERTScore to examine both the lexical and semantic quality of the model outputs. Through this comprehensive experimental design, we aim to illumi-

nate key factors that influence the effectiveness of LLMs in structured medical discourse, ultimately contributing valuable insights into the optimal configuration of models for real-world applications in the medical domain.

Generally speaking, the MARS dataset represents a valuable advancement in the study of medical dialogues, addressing challenges in both IE and online forum summarization. By building on cleaned forum data shared with DermaVQA [67], the dataset provides a solid foundation for analysis. It incorporates advanced annotations that go beyond conventional entity and relation extraction, capturing complex phrases related to treatment and diagnosis, thus offering a more nuanced understanding of medical discourse. Additionally, structured summaries crafted by trained annotators encapsulate key elements of medical discussions, including patient conditions, diagnoses, treatment plans, and recommendations. These features make the MARS dataset suited for evaluating large language models on medical tasks, introducing challenges that require advanced reasoning and the ability to interpret complex implications within dialogues. As a benchmark, MARS fills a gap in the existing literature, facilitating robust research and fostering the development of LLMs tailored to the medical domain. Ultimately, it aims to support the creation of more effective healthcare applications and research methodologies.

Following this introduction, Chapter 2 delves into related work, highlighting existing methodologies and datasets in online thread discussion summarization, with a focus on gaps in the medical domain. Chapter 3 presents the dataset composition in this study, detailing the data collection, annotation methods, summary composition techniques, and statistical analyses and inter-annotator agreement metrics to ensure the reliability of the annotations. In Chapter 4, we introduce experimental design for benchmarking language models. The experiments conducted are described, outlining the setup and presenting the results of the experimentation to assess the summarization models' performance. Chapter 5 discusses the implications of the findings, provides interpretations of the results, and suggests potential avenues for future research. Through this structure, the thesis aims to contribute valuable insights to the field of medical discourse summarization while addressing the existing

challenges in the area. Finally, we will conclude the entire thesis.

Chapter 2
RELATED WORK

In this chapter, we review related work, starting with research on online thread discussion summarization and existing datasets. Since our work involves not only building a dataset but also developing a summarization system, the third section provides an overview of existing summarization systems. Additionally, information extraction is an essential technique in our annotation process, dataset construction, and comparative experiments with summarization systems, so we summarize existing information extraction models in the fourth section. Finally, to evaluate our summarization models, it is necessary to explore evaluation metrics for natural language generation (NLG), which are covered in the fifth section.

2.1 Online thread discussion summarization

General domain Online thread discussion (also known as online forum discussion) helps share information and discuss problems of the same interest. Bhatia et al.[10] mention that thread discussions involve multiple users with various viewpoints and solutions. Hence, it is useful to generate a summary that gives users an understanding of the background of the entire discussion and an overview of different viewpoints. The format of online thread discussion is usually that a thread initiator posts a question and the other users reply and discuss [10, 38].

Even though online discussion threads have been popular for many years, studies on discussion thread summarization are limited given the large amount of data available. One of the main reasons is that discussion thread summarization datasets are hard to construct since various information is compacted in one thread. There is only a small extractive summarization dataset created in Bhatia et al.'s study [10]. Until more recently, some abstractive dis-

cussion thread summarization datasets have been proposed. For instance, **ConvoSumm** [20] is an abstractive conversation summarization benchmark containing news article comments, discussion forums and debates, community question answering, and email threads.

AnswerSumm [21] is another thread summarization dataset containing 4631 StackExchange question-answering discussion threads. Their annotation protocol has four steps: (1) answer sentence selection, (2) clustering, (3) cluster summarization, and (4) cluster summary fusion. All data are automatically segmented into sentences in the first place. In step (1), only relevant sentences (whether providing useful information to answering the user's question) are selected. Then, sentences sharing the same topics are grouped together irrespective of their polarities. Complex sentences might be clustered into multiple groups in step (2). There are no pre-defined number of clusters. In step (3), annotators are asked to summarize each cluster in one to four complete sentences. The summaries focus on the viewpoint presented in the clusters and try to include some details. In step (4), annotators combine cluster summaries in the previous step to make one coherent summary. Modifications like paraphrasing, and adding connectives are required to make the final summary more coherent.

Similar to the annotation methods in Fabbri et al.[21], Overbay et al.[50] present the first multimodal discussion summarization dataset, **M*R*EDDIT*S*UM**. Their annotation process has three steps. The first step is to summarize the original post. Annotators are asked to summarize the original post in a single sentence to show the intent of the original poster and the most relevant details from the image. The second step is to summarize the comment cluster. Before summarization, comments sharing a similar opinion are grouped by a RoBERTa-based fine-tuned model. Then, groups of comments as well as original posts and images are presented to annotators. They are asked to summarize the main opinions of each group within one or two sentences. To ensure consistency of summaries, commenters are all referred to as "Commenters" in contrast to "users", "people", or others. Annotators are also encouraged to relate details from images when necessary. The third step is to synthesize original post summarization and comment cluster summarization in descending order of their saliency scores. Then the synthesized summaries are presented to annotators again,

and frequency, as well as readability, are improved in this turn of summarization, including deleting repeated words, adding connectives, and rearrangement considering topics. All summaries are written in the past tense.

Medical domain Summarization in the medical domain has a few sub-fields, which include clinical notes [39], drug information [23], medical literature [41], and medical dialogue [30].

Medical-domain summarization has unique characteristics compared to domain-general or other domain-specific summarization tasks. Joshi et al. [30] observe that for medical dialogue summarization, it is important to (1) contain all medical conditions and terminology mentioned, (2) correctly discriminate all the negatives and affirmatives on medical conditions, and (3) attempt to copy from the source text but not be entirely extractive. They also observe that, unlike open-domain dialogues that involve long-term memory dependencies, there is an inherent local structure in patient history. Another big challenge in medical-domain summarization is the lack of large-scale annotated datasets [30].

Ben Abacha et al. [7] also mention in their doctor-patient conversation summarization dataset paper that even though the development of transformer-based language models has been benefiting the summarization of medical field, the lack of domain-specific and task-specific data and related evaluation metrics is still an issue. They also claim that summarizing doctor-patient conversations in a clinical setting involves unique challenges beyond standard natural language understanding and generation. Key considerations include the risk of omitting important medical information, which could affect patient outcomes, and the potential for inaccuracies or hallucinations that may impact clinical results. In Ben Abacha et al. [8], medical facts should include *problems*, *allergies*, *medical history*, *treatment*, *medications*, *test*, *laboratory/radiology results* and *diagnoses*.

2.2 Existing datasets on thread summarization

We conducted a comprehensive analysis of six existing datasets, examining them from several perspectives. Table 2.1 shows the differences between the existing dataset and the MARS

dataset. The details of each dataset are as follows.

Dataset Name	Summary Structure	Information Extraction	Domain Specific	Dataset Size
MTS-Dialog [7]	–	–	Medical	1,701
ACI-Bench [66]	*	–	Medical	67
MEDIQA 2021 [6]	–	–	Medical	–
mRedditSum [50]	*	*	General	–
ConvoSumm [20]	*	–	General	1,000
AnswerSumm [21]	*	–	General	4,631
MARS	*	*	Medical	998

Table 2.1: Comparison of existing datasets with our dataset MARS based on summary structure, use of information extraction, domain specificity, and dataset size. The domain column indicates whether the dataset is specific to a particular domain (i.e., medical) or general purpose. A * indicates a structured summary or the use of information extraction, while a cross (–) indicates the absence of these features.

MTS-Dialog The MTS-Dialog dataset [7][1] selected clinical notes from six most common types: general medicine, SOAP (Subjective, Objective, Assessment, Plan), neurology, orthopedic, dermatology, and allergy/immunology. The authors engaged eight medically trained annotators to develop dialogue summaries that effectively encapsulate the core elements of patient-physician conversations. These summaries are unstructured and evaluated using a range of metrics, including ROUGE-N, Fact Scores, BERTScore, and BLEURT. The annotation guidelines emphasize several key principles:

(1) Conversation Creation Rules dictate that conversations should be written in relation to the day of the patient's visit and framed within the context of either an outpatient or emergency room visit;

(2) Medical Terms Rules state that clinical notes may provide more detailed descriptions of problems, treatments, or tests than the conversational expressions used, such as how

[1]https://github.com/abachaa/MTS-Dialog

"Open reduction internal fixation (ORIF)" might be simplified to "We will have to do surgery on it";

(3) according to the Imaginary but Plausible Rule, if clinical notes are underspecified, plausible conversations may be created, though they should maintain a level of detail that surpasses that of the associated clinical notes, except for problems, treatments, and tests;

(4) Formatting Rules require annotators to adhere to standard transcription guidelines, including accurate phonetic representation, capitalization, and punctuation;

(5) the Conversation Characteristics' goal aims to capture a wide variety of dialogue types to simulate real doctor-patient interactions, incorporating natural speech disfluencies like false starts, filler words, interjections, interruptions, corrections, and the use of slang or colloquial terms.

The dataset comprises 1,701 pairs of dialogues and corresponding summaries. On average, the dialogues contain 9 turns, with a maximum of 103; they consist of an average of 11 sentences, with a maximum of 136, and contain an average of 142 words, extending up to 1,951 words. In contrast, the summaries average 3 sentences, reaching a maximum of 57. Several state-of-the-art transformer-based summarization models are used in the paper, including BART [35] and Pegasus [71], along with variants that were pre-finetuned on relevant datasets such as XSum [44] and Samsum [25], and incorporated augmented training data and guided summarization techniques.

ACI-Bench The ACI-Bench dataset [66][2] simulates three modes of clinical note generation from doctor-patient conversations: virtual assistant (virtassist), virtual scribe (virtscribe), and ambient clinical intelligence (aci). Transcripts were created by medical experts or ASR

[2]https://figshare.com/articles/dataset/aci-bench-corpus_zip/22494601

systems, while clinical notes were generated automatically and refined by domain experts. The original data included simulated doctor-patient conversations, realistic clinical notes augmented with imaginary EHR inputs, and ASR-generated transcripts that contained errors such as misheard terms or names. Through systematic processing, unsupported content (e.g., imaginary EHR inputs) was annotated and removed, ASR errors were corrected, and inconsistencies in clinical notes were resolved. To improve data utility, clinical notes were segmented into sections—subjective, objective_exam, objective_results, and assessment_and_plan—for granular evaluation. The dataset includes human-generated and ASR transcripts (both raw and corrected) alongside refined clinical notes, offering a high-quality benchmark for evaluating AI-assisted clinical note generation and evaluation.

These data exhibit a clearly defined structure, incorporating titles such as "CHIEF COMPLAINT," "MEDICAL HISTORY," and "HISTORY OF PRESENT ILLNESS," thereby reflecting a format akin to authentic medical records. The dataset is systematically divided into four sections: subjective, objective results, objective examination, and assessment and plan. Evaluation metrics employed include ROUGE (1/2/L), BERTScore, BLEURT, and a medical expertise score. While systematic annotation guidelines were established, specific details regarding these guidelines are not explicitly disclosed. The degree of agreement among annotators was assessed through partial span overlap, quantified by the F1 score, a standard statistical measure for evaluating binary classification accuracy. Statistical details of the dataset indicate that the training set comprises 67 encounters, with an average of 56 dialogue turns per entry, and the notes average 483 tokens in length with 48 sentences.

To benchmark their dataset, the study assesses several note-generation models, including (1) Transcript-Copy-and-Paste [25] that extract the longest and most relevant dialogue components; (2) Retrieval-Based Methods [31] that select relevant notes from the training corpus using UMLS concept set and document embedding similarity; (3) BART [35]-Based Models, including a version continued pre-trained on PubMed abstracts [34] and another fine-tuned on the SAMSum corpus [25], both with a 1,024-token limit; (4) LED-Based Models [5, 14], which use the Longformer-Encoder-Decoder architecture to manage longer tran-

scripts up to 16,000 tokens; and (5) OpenAI Models [3] like Text-davinci and GPT-4, which utilize specific prompts for structured clinical note summarization and implement rule-based post-processing for section detection. These models collectively evaluate the effectiveness of different summarization strategies within the corpus.

MEDIQA 2021 The MEDIQA 2021 dataset [6][4], integral to the BioNLP 2021 workshop, encompasses three subtasks: Consumer Health Question Summarization (QS), Multi-Answer Summarization (MAS), and Radiology Report Summarization (RRS). The QS subtask employs the MeQSum dataset, featuring 1,000 expert-annotated consumer health questions aimed at generating concise summaries that retain essential information. Evaluative measures, primarily ROUGE-2 alongside BERTScore and HOLMS, assess the quality of the generated summaries. The MAS subtask utilizes the MEDIQA-AnS dataset, aggregating multiple answers to medical questions while maintaining coherence in both extractive and abstractive formats. In the RRS task, the focus is on summarizing chest radiography reports from the MIMIC-CXR dataset, aiming to produce concise impressions based on detailed findings. Evaluation metrics include ROUGE-2 and Hamming similarity to ensure clinical validity. Baseline systems for each subtask incorporate state-of-the-art models, including PEGASUS [72] for QS and pointer-generator models for RRS, providing a robust foundation for participants' submissions. Overall, the MEDIQA 2021 tasks address critical challenges in automated medical summarization, facilitating advancements in the field.

MREDDITSUM The mRedditSum dataset [50][5] establishes a robust multimodal discussion summarization framework characterized by its summary structure that effectively synthesizes critical information from both text and images. Each summary consists of a one-sentence encapsulation of the original post's intent, derived from the accompanying image,

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

[4]https://github.com/abachaa/MEDIQA2021

[5]https://github.com/Koverbay/mredditsum

followed by condensed summaries of comment clusters that articulate diverse perspectives within the discussion. Evaluation metrics for mRedditSum include ROUGE scores—specifically ROUGE-1, ROUGE-2, and ROUGE-L—which measure n-gram overlap and the longest common subsequence between generated and reference summaries, alongside BERTScore which assesses semantic similarity through contextual embeddings.

Baseline systems for the dataset comprise a variety of extractive and text-only models. Extractive baselines include Lead-1, which utilizes the first sentence of the document; Lead-Comment, which summarizes using the top five comments from the thread; and Ext-Oracle, which extracts text passages to achieve the highest possible ROUGE score, representing optimal performance for extractive methods. Additionally, text-only baselines are evaluated, including GPT-3.5 [48], which leverages a large language model (text-davinci-003) for zero-shot summarization, as well as fine-tuned models like BART-base [35] and T5-base [53], renowned for their summarization efficacy. LongT5-base [26], an extension of T5 designed for longer sequences, is also included. These text-only models are pre-trained on the CNN/DailyMail dataset [43] and subsequently fine-tuned for the mRedditSum task, creating a comprehensive baseline for comparison. The annotation process, executed by qualified annotators following stringent guidelines, reflects high inter-annotator agreement levels, affirming the quality and reliability of the generated summaries.

ConvoSumm ConvoSumm [20][6] is a meticulously curated dataset designed to bridge the gap in abstractive summarization of diverse online conversations, encompassing formats such as news comments, discussion forums, community question answering platforms, and email threads. The dataset adopts a structured summary framework based on the issues–viewpoints–assertions model, enabling a nuanced capture of participants' perspectives. This structure mandates that summarizations distill key viewpoints while retaining specific assertions, thereby reflecting a comprehensive analysis of the dialogues. They select 500 discussion forums and debates from Reddit data in CoarseDiscourse [70]. There are another 500 in-

[6]https://github.com/Yale-LILY/ConvoSumm

stances from StackExchange(Stack) as the data for the community questionanswering subdomain. Their annotation protocol is inspired by [4] "issues–viewpoints–assertions" argument structures. Specifically, three key components of conversational dialogues are identified: *issues* (which people discuss), *viewpoints* (which people hold about the issues), and *assertions* (which people make to support their viewpoints). During the annotation process, crowdsourced workers are asked to summarize all viewpoints, and to the fullest extent include details from assertions and anecdotes. There are circumstances where many viewpoints are similar, suggested wording like "Most commenters suggest that..." and "Some commenters think that..." are provided for annotators. Apart from "issues–viewpoints–assertions" argument structures, there are overall summarization rules like (1) summaries need to be analyses but not responses, (2) summaries should be abstractive, and no more than five words in a row from the sources can be repeated, and (3) summary lengths should fall between 40 to 90 tokens.

The evaluation of summaries is anchored in robust metrics such as ROUGE scores, allowing for performance benchmarking against existing datasets. Notably, the reported percentage of novel n-grams surpasses that of the highly abstractive XSum dataset, indicating successful adherence to the abstractive approach as directed by the annotation guidelines. These guidelines are meticulously crafted, instructing annotators to produce paraphrased summaries with specified lengths, while also emphasizing the inclusion of diverse viewpoints.

Quality assurance mechanisms include thorough manual reviews and a feedback-centric workflow, fostering high inter-annotator agreement among a select group of experienced annotators. This rigorous process not only mitigates typical noise associated with crowdsourced data but also reinforces the reliability and coherence of the summaries generated. Consequently, ConvoSumm stands as a robust resource for advancing the state of conversation summarization, providing a validated framework for future research in this emerging domain.

AnswerSumm The AnswerSumm dataset [21][7] was created by annotators aiming to capture the essence of question-answer pairs through cluster summaries. These summaries have a structure within the dataset, where there are cluster summaries. It utilizes ROUGE (1/2/L), NLI (Natural Language Inference), and Semantic Area as its evaluation metrics. The annotation guidelines include Answer Sentence Selection, Clustering, Cluster Summarization, and Cluster Summary Fusion. The inter-annotator agreement is Fleiss Kappa value. The AnswerSumm dataset consists of 4,631 high-quality data points. Average statistics for the dataset include inputs with an average of 6.4 answers, 40.3 sentences, and 787 words; the output data shows an average of 2.6 clusters with summary lengths of 47 words.

2.3 Existing summarization systems

2.3.1 Extractive system

The SummaRuNNer system, designed by Nallapati et al. [42], significantly enhances extractive summarization through a multi-layered architecture that employs a bidirectional GRU-RNN. This design allows the model to capture sentence representations while accounting for both current context and previous sentences. The first layer processes each sentence sequentially, producing hidden states that inform a second layer RNN, which aggregates these states into a comprehensive document representation. A key innovation in SummaRuNNer is its interpretability, as it visualizes predictions based on features such as content, salience, and novelty. This transparency aids in understanding how each sentence contributes to the overall summary. Additionally, the system adopts an abstractive training approach, utilizing human-generated reference summaries to inform sentence selection without the need for traditional sentence-level labeling. By strategically selecting sentences based on their contribution to summary quality, SummaRuNNer optimizes performance, achieving results that are competitive with state-of-the-art methods in extractive summarization.

In their work, "Extractive Summarization of Long Documents by Combining Global and

[7]https://github.com/Alex-Fabbri/AnswerSumm

Local Context," Xiao and Carenini [64] propose a novel neural extractive summarization model specifically designed for long documents. The architecture integrates both global context—considering the entire document—and local context, focusing on the current topic segment to enhance summarization performance. The model sequentially processes sentences using a bi-directional GRU, effectively capturing contextual information from adjacent sentences. By leveraging this structure, the model demonstrates superior performance on datasets of scientific papers, such as Pubmed and arXiv, outpacing both extractive and abstractive models as evidenced by improvements in ROUGE-1, ROUGE-2, and METEOR scores. Remarkably, the ablation study reveals that the significant improvements stem primarily from effectively modeling local context, underscoring the design' s adaptability for longer documents. This approach achieves a more nuanced understanding of sentence relevance, contributing to more coherent and focused summaries.

Song et al. [59] explored medical conversation summarization using a Chinese corpus gathered from an online healthcare platform. They developed a hierarchical encoder-tagger model (HET) to perform extractive summarization, specifically focusing on two types of summaries: problem statements and treatment recommendations. Their experiments showed that HET outperformed strong baseline models, and incorporating conversation-related features further enhanced its performance, enabling the generation of high-quality summaries from medical dialogues.

Zhong et al. [77] propose a transformative approach to neural extractive summarization by framing the task as a semantic text matching problem. Instead of treating sentence extraction as an isolated process, the MATCHSUM framework matches candidate summaries —with semantic representations derived from a source document—within a shared semantic space. This shift enables better evaluation of summaries based on their overall semantic similarity to the document, rather than solely on individual sentences. The authors utilize a Siamese-BERT architecture, leveraging the strengths of pre-trained BERT to create meaningful embeddings for both the document and candidate summaries. By employing cosine similarity to assess match quality, their model achieves significant improvements in summa-

rization effectiveness, notably reaching a new state-of-the-art ROUGE-1 score of 44.41 on the CNN/DailyMail dataset. Comprehensive experiments across various benchmark datasets further validate the superiority of their framework, demonstrating its potential within the extractive summarization field and indicating that the capabilities of this matching-based approach have yet to be fully realized.

2.3.2 Abstractive system

Enarvi et al. [19] investigated the automatic generation of medical reports from ASR-derived transcripts of patient-doctor conversations using both RNN and Transformer-based sequence-to-sequence models. They enhanced these architectures with techniques like the pointer-generator network for copying conversation segments and a hierarchical RNN encoder to speed up training. Their analysis, conducted on a dataset of 800k orthopedic encounters, demonstrated that Transformer models outperformed RNNs in accuracy while requiring less training time. These findings suggest that sequence-to-sequence modeling is a promising method for large-scale automatic medical report generation.

PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) [72] is a state-of-the-art transformer model specifically designed for abstractive text summarization. This model diverges from traditional pre-training approaches by introducing a novel objective called Gap Sentence Generation (GSG), where key sentences are masked in the input text and subsequently generated as a unified output sequence. This mirrors the process of creating summary-like text, allowing PEGASUS to capture the essence of the document effectively. The model leverages large text corpora, including C4 and HugeNews, for its pre-training phase. It employs various gap sentence selection strategies—Random, Lead, and Principal—where "Principal" selection, based on ROUGE scores, proved to be the most effective. PEGASUS was evaluated across 12 diverse downstream summarization tasks, spanning domains from news articles to scientific publications, achieving remarkable results that consistently outperformed existing models based on ROUGE metrics. Notably, it demonstrated exceptional performance even in low-resource scenarios, achieving state-of-the-art results on

datasets with as few as 1,000 examples. Human evaluations further validated its output quality, showing that the summaries generated by PEGASUS align closely with human-level standards across multiple datasets. Overall, PEGASUS represents a significant advancement in the field of abstractive summarization, combining innovative pre-training strategies with robust performance across various applications.

The BRIO (Bringing Order to Abstractive Summarization) method [37] for abstractive summarization introduces an innovative training approach that diverges from traditional maximum likelihood estimation (MLE). In conventional MLE frameworks, models are trained with the assumption of a deterministic distribution, where all probability weight is allocated exclusively to the reference summary. This can hinder performance when the model must evaluate multiple system-generated summaries that differ from the reference. In contrast, BRIO proposes a non-deterministic paradigm where candidate summaries are assigned probability mass based on their assessed quality. This facilitates a more effective comparison during inference. By incorporating a novel contrastive loss, BRIO aligns predicted probabilities of candidate summaries with actual quality metrics, enhancing the model's dual functionality as both a generator of summaries and an evaluator of their quality. Notably, BRIO has achieved state-of-the-art results on established benchmarks, such as 47.78 ROUGE-1 on the CNN/DailyMail dataset [28] and 49.07 ROUGE-1 on XSUM [44]. This work underscores the model's ability to generate summaries and accurately estimate their quality, highlighting a significant advancement in the field of abstractive summarization.

Krishna et al. [32] introduced Cluster2Sent, a hybrid algorithm for generating SOAP notes from doctor-patient conversations. This approach combines both extractive and abstractive methods: it first extracts key utterances for each section of the summary, then clusters these related utterances, and finally creates a single summary sentence for each cluster. This model significantly outperformed purely abstractive methods in generating factual and coherent summaries, as verified by expert evaluations, and demonstrated similar advantages when applied to the publicly available AMI dataset.

2.3.3 LLM summarization

Given the fast development of Large Language Models , various NLP tasks (e.g., information extraction, classification, summarization) can be directly conducted. Many LLMs have shown promising results across domains in zero-shot and few-shot tasks.

Zhang et al. [75] analyze the effectiveness of LLMs in automatic summarization through a comprehensive human evaluation involving ten models specifically on the news summarization task. They compare ten large language models across different pretraining methods, prompts, and model scales. The models evaluated include various versions of OpenAI's GPT-3, specifically zero-shot and instruction-tuned variants, such as Instruct Curie and Instruct Davinci [13], and InstructGPT [48]. OPT 175B [73], GLM [18], Cohere xlarge v20220609 [15], and Anthropic-LM v4-s3 [2] are also compared in the study. Additionally, the study benchmarks state-of-the-art fine-tuned models, notably PEGASUS [72] and BRIO [37]. The evaluation focuses on summarization tasks using the CNN/DailyMail [28] and XSUM [44] datasets, allowing for insights into the performance of these LLMs versus traditional fine-tuned models across multiple metrics related to summarization quality. Three aspects are manually evaluated to compare the gold-standard and generated summaries, which include *faithfulness*, *coherence*, and *relevance*. Six automatic metrics are also implemented including ROUGE-L, METEOR, BertScore, BLEURT, and BARTScore.

Their study reveals that instruction tuning, rather than model size, plays a pivotal role in enhancing the zero-shot summarization capabilities of LLMs. Furthermore, they highlight the detrimental impact of low-quality reference summaries on prior evaluations, which likely understate human performance and diminish the perceived effectiveness of both few-shot and fine-tuning techniques. By utilizing high-quality summaries sourced from freelance writers, the authors demonstrate that LLM-generated summaries perform comparably to those produced by humans, despite notable stylistic differences. Their insights underscore the need for improved reference quality in summary evaluations to ensure accurate benchmarking across model types and configurations, suggesting that current benchmarks may yield limited value

if reference quality issues are not addressed.

2.4 Existing information extraction models

2.4.1 Traditional NLP methods

Zhang et al. [76] proposed a Medical Information Extractor (MIE) to facilitate the automatic conversion of medical dialogues into Electronic Medical Records (EMRs), aiming to alleviate the burdens of physicians who find EMR documentation tedious. They employed a window-sliding annotation method for labeling online medical consultation dialogues, which simplifies the process compared to sequential labeling. MIE is designed to extract vital information such as symptoms, surgeries, tests, and their statuses by utilizing a deep matching architecture that considers interaction during dialogue turns. Experimental results indicate that MIE effectively extracts medical information from doctor-patient conversations.

2.4.2 LLM methods

Due to the specificity of datasets in the medical field [63], clinical information extraction faces numerous challenges. Currently, it can mainly be approached in two ways: the first method involves using pretrained language models for clinical NLP, while the second method uses large language models (LLMs, hereafter referred to as large models) based on prompt-based learning [1].

According to the study by Wu et al. (2020) [63], RNNs combined with word2vec embeddings are the most commonly used methods in medical natural language processing tasks, with the majority of these tasks being information extraction tasks. With the development of BERT, some domain-specific pretrained language models such as ClinicalBERT, SciBERT, BioBERT, and PubMedBERT have also been introduced.

With the emergence of large language models like GPT-3, prompt-based learning (or in-context learning) has also been applied in the medical domain [1, 27, 40], achieving better results in zero-shot or few-shot information extraction tasks. However, this method requires

more intricate designs during execution to obtain optimal and precise results [1]. In addition to prompt-based learning, Chain of Thought (CoT) [62] and self-consistency [61] have also been utilized to leverage instructions for completing medical natural language processing tasks [45].

Agrawal et al. (2022) [1] proposed that large language models like GPT-3 [12] can achieve good results in zero-shot or few-shot information extraction in the medical field. They tested GPT-3 [12] and InstructGPT [49] on five clinical information extraction tasks, including clinical sense disambiguation, biomedical evidence extraction, coreference resolution, medication status extraction, and medication attribute extraction. They designed corresponding instructions based on five samples selected from the validation set of each of these five tasks. To enhance the extraction performance of large language models, they proposed a "resolver" mechanism, which processes the outputs from the large model and converts them into the final results. The resolver is designed according to the characteristics of each task, aiming to make the large model's complex output more structured. They found that in their study of clinical extraction tasks, the GPT-3 system significantly outperformed existing zero-shot and few-shot baselines.

Nori et al. (2023) [45] further discovered that innovative instructions can unlock deeper expert capabilities of large language models. They demonstrated that under the influence of their designed MedPrompt (a combination of several instruction strategies), GPT-4 [47] performs best on the benchmark dataset of MultiMedQA [57], exceeding the performance of state-of-the-art expert models such as Med-PaLM 2 [58], while reducing the number of model invocations by an order of magnitude. MedPrompt mainly includes three techniques: dynamic few-shot selection, self-generated CoT, and choice shuffle ensembling. Figure 2.1 illustrates the components and effects of MedPrompt [45].

First, dynamic few-shot selection focuses on the quality of the samples used for training. Since the small sample examples used for prompting specific tasks are typically fixed, they must be broadly representative in the testing examples. Apart from having domain experts select suitable samples, the authors measured the semantic distances between each

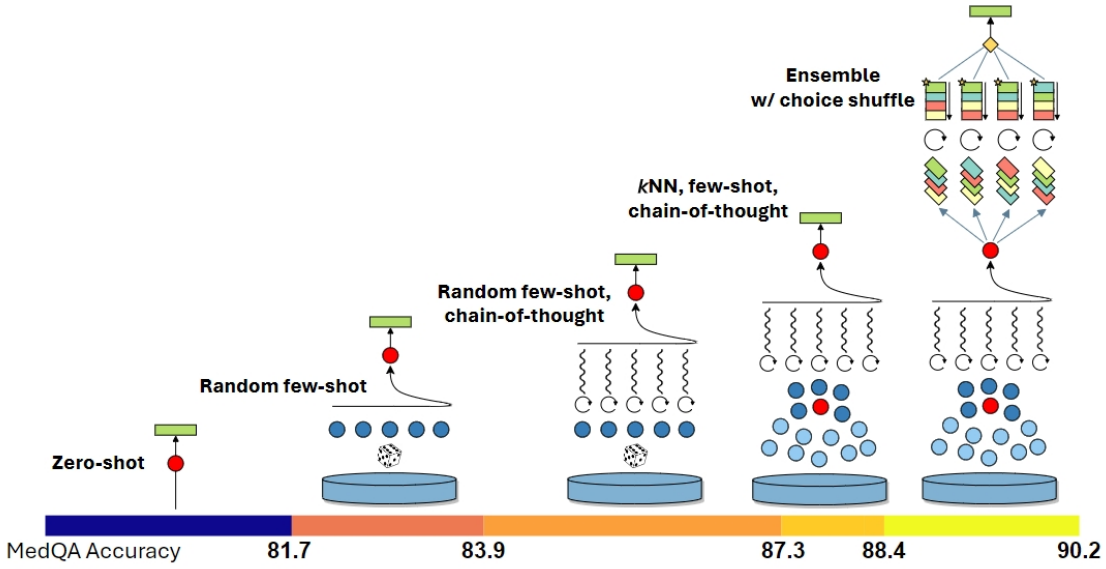

Figure 2.1: Visual illustration of Medprompt components and their additive contributions to performance on the MedQA benchmark. The prompting strategy integrates kNN-based few-shot example selection, GPT-4–generated chain-of-thought prompting, and answer-choice shuffled ensembling. The relative contributions of each component are depicted at the bottom. Adapted from [45] Figure 4.

testing sample and all samples in the training set through k-NN clustering. Specifically, they first used text-embedding-ada-002 to obtain vector representations of the training and testing questions. Then, for each testing question x, they retrieved its k-nearest neighbors $x_1, x_2, ..., x_k$ from the training set. Given a predefined similarity metric d, such as cosine similarity, the neighbors were sorted with $d(x_i, x) \leq d(x_j, x)$ for $i < j$, and the closest training questions were selected as small samples.

For self-generated CoT, the authors first compared the CoTs generated by experts with those generated by GPT-4 on the testing set questions, discovering that GPT-4 could generate detailed explanations suitable for small sample CoT demonstrations. Therefore, they instructed the model to self-generate CoTs based on the training set. However, self-generated CoTs carry the risk of containing hallucinations or incorrect CoTs. To mitigate the impact of such risks, they instructed GPT-4 to generate answers that are most likely to follow the CoTs. If the generated answer does not match the true answer, they completely discard the sample, assuming that the CoT cannot be trusted. Although hallucinations or incorrect CoTs can still produce the correct final answers, they found that the aforementioned approach was effective for filtering. They also observed that the CoTs generated by GPT-4 are longer and provide more detailed step-by-step reasoning logic than those handwritten by experts. Recent work [22, 65] has also found that foundational models can outperform expert-written instructions.

Since GPT-4 exhibited a preference for certain options in multiple-choice settings, they employed option shuffling and used self-consistency prompting to check the consistency of answers across different orderings. This approach increases the diversity of reasoning paths, thereby improving the quality of the final ensemble. They also applied this technique during the intermediate CoT generation process for generating training examples. For each example, they shuffled the choices multiple times and generated a CoT for each variant, ultimately retaining only the examples with the correct answers.

2.5 NLG evaluation metrics

To give a reliable evaluation in Natural Language Generation (NLG) is challenging, mainly due to the high cost of human-expert evaluation and the complexity of human language [8]. Metrics commonly used in NLG evaluation belong to two categories: N-gram-based metrics (e.g., ROUGE [36], BLEU [51]) and embedding-based metrics (e.g.BertScore [74], SUPERT [24]). Both measure the similarity between generated texts and reference texts.

There are other domain-specific automatic evaluation metrics proposed but few focus on the clinical field, where medical facts omission can largely affect the performances [8]. They proposed four types of evaluation metrics for the task of clinical note generation[8], and compare these metrics from different criteria such as factual correctness, hallucination, and omission rates.

The first type of metric is a Knowledge Graph Embedding (KGE)-based metric. They proposed a model called MIST, which uses knowledge graphs to provide additional semantic information in the embeddings. The MIST focuses on the similarities of *medical concepts* between the generated texts and reference texts. The MIST value is calculated as follows:

$$MIST(S, R) = \frac{1}{\mid R \mid} \sum_{c \in S} \max_{r \in R} cos(G_c, G_r) \tag{2.1}$$

G_c represents the graph embeddings of each concept c, which belongs to the set of concept S from the system-generated summaries. G_r is the graph embeddings of each concept r, which belongs to the set of concept R from the reference summaries. Cosine similarities between all concepts in generated texts and their closest concepts in reference texts are summed up, and the sum is divided by the count of reference concepts, which formulates the recall-oriented MIST value.

The second type of metrics is Finetuning-based Metrics. Based on the BLEURT-512 model [55], [8] finetune the model with a total of 6367 family medicine and orthopedic encounters, and present the ClinicalBLEURT.

---

[8]https://github.com/abachaa/EvaluationMetrics-ACL23/tree/main

The third type of metric is Customized Model-based Metrics, which assign a higher weight to medical terms. The modified the BARTScore [69] and the BertScore [74] in the following way:

$$MedBARTScore = \sum_{i=1}^{m} w_t logp(y_t \mid y < t, x) \tag{2.2}$$

where x is the source sequence, and $y = (y\_1, ..., y\_m)$ are the tokens of the target sequence, which has a length of m.

$$MedBERTScoreP = \frac{1}{\mid \hat{x} \mid} \sum_{\hat{x_i} \in \hat{x}} w_x \max_{x_j} x_i^T \hat{x_j} \tag{2.3}$$

where x is the reference summary and $\hat{x}$ is the candidate summary.

For both metrics, for non-medical words $w = 1$, and for medical words $w = 1 + \alpha$. Ben Abacha et al. [8] test α from $[0.1, 1.5]$ and find the best α is 1.0.

Additionally, to deal with the disadvantage of traditional evaluation metrics like ROUGE, where only texts under encode-limit of the pre-trained models can be encoded, Ben Abacha et al. [8] propose the Sliding Window Policy. However, for our project, the summaries are not likely to exceed 512 tokens. We will leave the policy for future reference.

The fourth type focuses on Ensemble Metrics, in which multiple metrics are combined. Ben Abacha et al. [8] select $C_1 = \{MIST, ROUGE - 1 - R, BERTScore\}$ and $C_2 = \{MIST, ROUGE - 1 - R, BLEURT\}$. $Z_m(x)$ is the normalized Z_{score} of a metric m, μ_m is the mean value of m over the summaries set, and σ_m is the standard derivation of m. The formulas are:

$$Z_m(x) = \frac{x - \mu_m}{\sigma_m} \tag{2.4}$$

$$MIST_{Comb1}(x) = \frac{1}{3} \sum_{m \in C_1} Z_m(m(x)) \tag{2.5}$$

$$MIST_{Comb2}(x) = \frac{1}{3} \sum_{m \in C_2} Z_m(m(x)) \tag{2.6}$$

Ben Abacha et al. [8] conclude that metrics good at capturing factual accuracy are not necessarily good at capturing hallucination and key medical fact omission. Their language-model-based and ensemble metrics can outperform SOTA N-gram metrics like ROUGE on unbiased reference summaries.

Similar to Ben Abacha et al. [8], Yim et al. [68] use four automatic evaluation metrics and their benchmarking automatic visit note generation. The metrics include N-gram metrics (ROUGE 1/2/-L), an embedding-based metric (BERTScore), a model-based metric (BLEURT), and a medical-concept-based metric (MEDCON) to judge the accuracy and consistency of clinical concept.

Chapter 3

DATASET CREATION

In Chapter 3, we primarily explain the construction of the MARS dataset. We will elaborate from the perspectives of dataset construction methodology, data collection and cleaning, data annotation, summary composition, and data statistics.

3.1 Methodology

The methodology for constructing the MARS dataset involves several key steps, ensuring both the quality and relevance of the data.

Firstly, data is collected from the platform, where it is sourced from relevant threads or discussions that align with the objectives of the dataset. This collection process is designed to ensure diversity and comprehensive representation of the types of data needed for further annotation and analysis.

Next, the collected data undergoes a cleaning process to remove any irrelevant or extraneous information, ensuring that the dataset maintains its integrity and focus (see section 3.2). This step is crucial to eliminate noise and prepare the data for the next stages.

Following data cleaning, the dataset proceeds to manual Information Extraction annotation, which is thoroughly explained in Section 3.3. During this stage, trained annotators manually extract key information from the data, categorizing it based on relevant attributes. This process ensures that the data is well-structured and provides the foundation for subsequent tasks.

Once the data is annotated, an automatic summary generation step is performed, which is further detailed in Section 3.4. In this step, machine learning models are used to generate structured summaries based on the annotated data, transforming the raw content into concise

and coherent summaries that capture the essential information.

To illustrate these processes, we can refer to Figure 1.1, which shows the processed data collected from the platform, Table 1.1, which presents the annotated results from the manual IE annotation, and Figure 1.2, which showcases the final structured summary generated from the data.

3.2 Data collection and cleaning

Data Source The MARS dataset is based on the same forum discussion data with DermaVQA [67], which is from IIYI (爱爱医). IIYI.com is an online platform in China where users can consult medical professionals by submitting their questions, often accompanied by relevant files and images. Responses may come from multiple contributors. User profiles display verification badges, such as medical license credentials, and rankings based on upvotes from previous interactions. Primarily serving an educational purpose, the platform is used by both patients and doctors. For their study, they collected discussion threads from the forum focused on dermatology and sexually transmitted diseases (皮肤及性传播疾病讨论版). The specific post IDs are available in our published dataset. DermaVQA contains multimodal data including images and thread texts. However, to make our annotation task more standard and focused, we only use the text data. [1]

Data Cleaning In the IIYI dataset, responses lacking substantive value (e.g., comments like "I also want to know the answer") were excluded. Additionally, in the DermaVQA dataset, threads were filtered out if they contained no images, no responses, or lacked any meaningful contributions. Our dataset is annotated based on the cleaned data. During the annotation process, we noticed that a very small number of words were masked by the "**" symbol (possibly due to being sensitive terms on the internet). However, this had no significant impact on text comprehension or annotation quality. Therefore, if "**"

[1]Even though our dataset does not make use of the images, we believe when researchers benchmarking summarization or information extraction systems with MARS, adding images may help by adding more contexts and knowledge.

appears within the segments to be extracted, we will include it directly without additional processing. [2]

3.3 Data annotation

3.3.1 Annotation guidelines

We outline the annotation process for our dataset using the BRAT annotation tool, focusing on the specific labels employed and the accompanying guidelines. The annotation framework is designed to ensure that relevant information is captured systematically, facilitating a clearer understanding of the data.

The first set of labels pertains to patient demographics and clinical background. The label **Age** captures the patient's age, including expressions like "xxx-year-old," "baby," and "elderly." In instances where both a specific age (e.g., "30-year-old") and a more general term like "elderly" are present, annotators should prioritize the specific age in the annotation. Similarly, the **Sex** label reflects the sex of the patient and encompasses terms beyond just "male" and "female," including familial terms like "husband" or "sister." If both a generic and specific term appear, the specific term (e.g., "female") should be highlighted. The **Clinical_History** label denotes the patient's history of prior medical conditions, which, although not directly related to the current problem, can provide valuable context for diagnosis.

Another critical aspect of the annotation process involves the identification of current medical concerns. The **Problem_Current** label represents the primary request made in the post, sometimes consisting of multiple fragments. Annotators must assess whether each piece of information is essential, asking themselves if omitting any detail would lead to a loss of key points. The **Duration** label indicates how long the identified problem has been present and should explicitly link to the Problem_Current to clarify the timeline and

[2] In the future, when enhancing the dataset and constructing multilingual versions, we may consider manually retrieving the original text to replace "**" with the original wording. However, due to current resource limitations, no such action will be taken at this stage.

context. On BRAT, **Problem_Current** and **Duration** will be linked together if the **Duration** refers to the duration of the **Problem_Current**.

For historical information, the **Previous_Diagnosis** label captures any earlier diagnoses related to the current problem, helping distinguish relevant past conditions from historical notes. It is essential to focus on the disease itself, avoiding broader phrases that dilute the specificity of the annotation. The **Previous_Treatment** label documents any treatments the patient has received concerning the current concern, which may include surgeries or medication regimens.

The final set of labels pertains to the commenters' contributions. The **Diagnosis** label captures the diagnosis offered by the commenter, allowing for various semantic values including *"present, negated, hypothetical_present, hypothetical_negated, possible, conditional*". This label emphasizes clarity in capturing hypothetical scenarios, possibilities, and conditional statements. Note that "hypothetical" means what is implied but not directly stated. For instance, For example, if a poster describes a symptom in their post and a commenter states, "The symptoms of Disease A are...," where the symptoms mentioned by the commenter differ from or even contradict those described by the poster, the commenter is implying "not Disease A" without explicitly stating it. In such cases, we assign the value "hypothetical_negated" to the diagnosis "Disease A." Additionally, the **Test** label includes any suggested diagnostic tests, which must be linked back to the appropriate diagnosis, while the **Treatment** label captures any proposed treatment options, ensuring these are also connected to the relevant diagnosis.

Several important rules govern the annotation process to enhance clarity and consistency. For repetitive content, only the first instance should be highlighted to avoid redundancy and ensure clear communication. When dealing with similar pieces of information, annotators should focus on highlighting the more specific terms, promoting nuanced understanding. Furthermore, if the content is fragmented across multiple lines, annotators should utilize the "add fragment" feature in BRAT rather than attempting to highlight across lines; this practice maintains the context and structural integrity of the information. The following

tables (Table 3.1, 3.2, 3.3, 3.3.1) show a clear breakdown of the annotation guidelines.

Label Name	**Description**	**Example**
Age	Patient' s age, not only includes "xxx-year-old" , "baby, elderly, etc." can also be included. However, if "xxx-year-old" and "elderly" both exist, go for the more specific one (xxx-year-old).	18 岁，老人，婴儿...
Sex	Patient' s sex, not only includes "male/female" , "husband, sister, etc." which can indicate the sex of the patient should be highlighted. However, if "female" and "the old lady" both exist, go for the more specific one (female).	女，姑娘，...
Clinical_History	The clinical history of the patient. This might not be related to the Problem_Current, but it shows the patient' s past disease which may help the diagnosis of the Problem_Current.	

Table 3.1: The labeling scheme for the poster content: focusing on patient profile information such as age, sex, clinical history, and other relevant demographic and medical details.

3.3.2 Annotation setup

We use BRAT[3] as our annotation tool. Based on the previously cleaned data, we transform the JSON data files into TXT files which can be used in BRAT (as an example shown in

[3]https://brat.nlplab.org/

Label Name	Description	Example
Problem_Current	The main request from the poster. Sometimes, it consists of more than one fragment. When judging whether the content should be highlighted, annotators can think "If this piece of information is not included in the summary, will it lose key points?"	求诊断，...
Duration	The duration of the Problem_Current. Please do not forget to link the Problem_Current and Duration when annotating in brat.	二十年，昨天起，...

Table 3.2: The labeling scheme for the poster content: focusing on main issues of the post and corresponding duration

Figure 3.1). If posters reply in the comment section, their author_id will be highlighted on the interface (shown in Figure 3.2). Two native Chinese-speaking annotators (one with a medical and one with a clinical natural language processing background) have been trained to use BRAT and to annotate the data under the following annotation guidelines.

Label Name	Description	Example
Previous_Diagnosis	Previous diagnosis for the Problem_Current. Please pay attention to distinguishing Previous_Diagnosis and Clinical_History. If a certain diagnosis exists and does not belong to Problem_Current, it may be the patient' s clinical history. When highlighting diagnosis, try to only highlight the disease name. For example, instead of highlighting "is disease A" , you only need to highlight "disease A." However, the modifier of the disease cannot be neglected, such as "serious disease A."	
Previous_Treatment	Previous treatment for the Problem_Current. The treatment can be surgery, medication, etc.	

Table 3.3: The labeling scheme for the poster content: focusing on precious diagnosis and treatment for current issues

3.3.3 Inter-Annotator Agreement

The inter-annotator agreement (IAA) is assessed through a methodical approach that includes both label count difference and F1 score calculation.

Firstly, the label count difference ("label_num_diff") is calculated by comparing the counts of each label between two annotation files, with "label_total" representing the average total number of annotations for each annotator. The "label_num_diff_ratio" is derived from the ratio of these count differences to the total number of labels, providing a quantifiable

Figure 3.1: An example of the BRAT annotation tool interface

measure of disagreement. The results are shown in Table 3.5.

In conjunction with label count analysis, the F1 score for each label is computed by matching annotations based on their span positions (span_start and span_end). Annotations are deemed a match if their spans overlap, leading to specific classifications: True Positives (TP) for correctly matched annotations, False Positives (FP) for annotations present in annotator-2' s output but absent in annotator-1's, and False Negatives (FN) for annotations found in annotator-1' s annotations but not in annotator-2' s output. The F1 score is then calculated based on these counts, reflecting the accuracy of the annotations.

Additionally, certain labels are treated with particular rules to enhance the robustness of the comparison. For**Problem_Current** annotations with the same turn_id are merged prior to analysis to account for contextual relevance. For "Diagnosis," the evaluation not only considers span positions but also checks for content accuracy in the Val field (e.g., distinguishing between present and negated cases). This comprehensive framework facilitates

Figure 3.2: An example of a poster's ID is highlighted in the comment section on the BRAT interface

a nuanced understanding of annotation agreement and discrepancies, contributing to the overall reliability of the annotation process (shown in Table 3.6.)

For the training dataset, we observe a mean IAA of 1.00 in **Age** and **Sex** extraction, indicating perfect agreement among annotators for these categories. However, metrics related to clinical conditions exhibit more variability, with mean F1 scores for **Problem_Current** at 0.64, **Diagnosis_Val** at 0.62, and **Previous_Diagnosis** at 0.92, showing a spectrum of agreement that suggests certain complexities or ambiguities in the annotation of clinical data. This may be because **Problem_Current** is not a common or clearly defined concept; it refers to the primary concern of the poster. Sometimes it even overlaps with **Previous_Diagnosis** and **Previous_Treatment**, which adds complexity to the annotation process. Moreover, because the number of documents in the training set is significantly larger than that in the validation and testing datasets, and there are greater differences in document length and annotation difficulty, the scores for the training set may be relatively lower compared to those for the validation and testing datasets.

In the validation dataset, similar trends are apparent, with **Age** and **Duration**

classifications achieving perfect agreement (F1 scores of 1.00). The **Problem_Current** metric shows an increase in agreement to 0.70, while **Diagnosis_Val** also shows a slight uptick to 0.66, hinting at improved clarity or a more straightforward categorization task in validation. The **Previous_Diagnosis** and **Previous_Treatment** remain consistent with scores of 0.83 and 0.90, respectively.

The testing dataset presents a marginally higher overall agreement profile, with the mean F1 scores for **Age**, **Sex** and **Duration** maintaining a perfect 1.00. A notable increase is observed in the **Problem_Current** metric, climbing to 0.74, indicating a potential enhancement in the consistency of this particular annotation. Overall, while certain categories demonstrate high levels of agreement across datasets, discrepancies remain in others, particularly those related to clinical history and diagnosis, which may be influenced by the inherent complexities in these areas. The overall consistency in IAA across the three datasets underscores the reliability of the annotation process, while also identifying specific areas in need of refinement and clearer definition.

3.4 Summary generation

After gaining the annotations from two annotators and resolving the major differences, extractive summaries can be composed automatically based on the annotations (i.e., information extracted from the thread discussions.) Here is a sample of how the summary looks like:

【病人信息】： 无,无
【病史】： 无
【前期诊断】： 无
【前期治疗】： 达克能宁喷雾两个月无明显效果,医生指导用鸡眼膏，之后出现变红变多
【发帖人主要诉求】： 帮忙诊断一下,三个月前出现,自己用达克能宁喷雾两个月无明显效果，医生指导用鸡眼膏，之后出现变红变多
【持续时间】： 三个月

【评论区诊断】：
7人认为是跖疣
3人认为可能是跖疣
2人认为不是鸡眼
1人认为是鸡眼

【治疗方案】：
激光治疗（基于诊断：是跖疣）

【检查建议】：
无

First, **Patient Information** (病人信息) uses the labels **Sex** and **Age**. The summarization rule requires content from these labels to be concatenated and sorted by the 'span_start' field. The output format is as follows: 【Patient Info (病人信息)】: Gender, Age. If no data is available, it outputs "None" (无).

Second, **Medical History** (病史) corresponds to the **Clinical_History** label. Here, content is extracted and concatenated under this label, formatted as: 【Medical History (病史)】: Clinical history content. If no data is available, the output is "None" (无).

Next, **Prior Diagnosis** (前期诊断) is based on the **Previous_Diagnosis** label. The content of this label is extracted and concatenated, with the output formatted as follows: 【Prior Diagnosis (前期诊断)】: Diagnosis content. If there is no data, the output defaults to "None" (无).

For **Prior Treatment** (前期治疗), the **Previous_Treatment** label is used. Extracted and concatenated information is presented as: 【Prior Treatment (前期治疗)】: Treatment content. If no data is available, the output will be "None" (无).

The **Main Problem** (发帖人主要诉求) label, **Problem_Current**, is used to gather

content describing the main issue, which is extracted, concatenated, and formatted as follows: 【Main Problem (发帖人主要诉求)】: Problem content. In cases where no data is present, "None" (无) is displayed.

Similarly, **Duration** (持续时间) is summarized from the **Duration** label, where extracted content is formatted as: 【Duration (持续时间)】: Duration content. When no data is available, the output is "None" (无).

Diagnosis from Comments (评论区诊断) employs the **Diagnosis** label. Here, the diagnosis-related data is extracted from comments, with occurrences counted based on the 'content' and 'Val' fields. The 'Val' field maps to descriptive terms such as "is" (是), "is not" (不是), or "possible" (可能是). The output is structured as 【Diagnosis from Comments (评论区诊断)】: <Count> people believe <is/possible/is not> <Diagnosis content>. If no data is available, the output will be "None" (无).

In the case of **Treatment Plan** (治疗方案), the **Treatment** label is used. Treatment-related content is extracted, and if linked to a specific diagnosis, that diagnosis information is included in the output. The format is as follows: <Treatment content> (Based on diagnosis: <Diagnosis content>). If no diagnosis linkage is present, only the treatment content is output. Absence of data results in "None" (无).

Finally, **Test Suggestions** (检查建议) are summarized from the **Test** label. Here, test-related content is extracted, and if linked to a specific diagnosis, the diagnosis information is incorporated. The output format is as follows: <Test content> (Based on diagnosis: <Diagnosis content>). If no linked diagnosis is available, only the test content is displayed. If no data is available, "None" (无) will be shown.

In this manner, the structured process enables consistency in summarization, ensuring that missing elements are handled systematically, which aids in standardized clinical analysis for research purposes.

3.5 Data statistics

The dataset split remains the same as the split of DermaVQA [67] shown in Table 3.7. The table provides an overview of the dataset split across the training, validation, and test sets, including the number of annotations and the distribution of instances across various categories. The training set is the largest, containing 842 annotations and a comprehensive range of categories, with 5,505 instances of diagnoses and significant representation across other categories such as treatments, current problems, and clinical history. The validation set, with 56 annotations, serves to fine-tune the model, featuring 404 diagnosis instances alongside other data points. Finally, the test set includes 100 annotations, designed to evaluate model performance, with a balanced distribution of instances across all categories. These statistics highlight the dataset's structure and its suitability for both training and evaluation purposes.

Label Name	Description	Example
Diagnosis	Diagnosis given by the commenter. Values should be given based on the semantics: - Present: 是 - Negated: 不是、不像... - Hypothetical_present：not literally stated but implied as present - Hypothetical_negated：not literally stated but implied as negated - Possible: 可能、应该、或... - Conditional: 如果... 就是...	
Test	Suggested test given. If the test suggestion is given based on a certain diagnosis, it should be linked to the Diagnosis.	
Treatment	Suggested treatment given (medication, surgery, ⋯). If the treatment suggestion is given based on a certain diagnosis, it should be linked to the Diagnosis.	

Table 3.4: The labeling scheme for the content of the comments: including the *diagnoses* provided by the commenters with attribute values that represent the certainty of the diagnosis (present, negated, hypothetical_present, hypothetical_negated, possible, conditional), and (corresponding) *treatment* and *test* suggestions.

Column	Train	Valid	Test
annotation1_label_total	15.05	15.75	17.64
annotation2_label_total	15.05	13.44	15.86
label_num_diff	2.69	2.80	2.77
label_num_diff_ratio	0.18	0.18	0.17

Table 3.5: IAA on average label counts per sample for train, validation, and test sets. The label_num_diff and label_num_diff_ratio metrics quantify the disagreement between two annotators.

Column	Train	Valid	Test
age	1.00	1.00	1.00
sex	1.00	0.98	1.00
problem_current	0.64	0.70	0.74
duration	0.98	1.00	1.00
previous_diagnosis	0.92	0.83	0.95
previous_treatment	0.89	0.90	0.93
clinical_history	0.94	0.92	1.00
diagnosis	0.94	0.94	0.96
test	0.96	0.98	0.92
treatment	0.90	0.96	0.94
diagnosis_val	0.62	0.66	0.68

Table 3.6: Average F1 scores in IAA for label annotation per sample across training, validation, and test sets. See above for calculation details. The results presented are the mean of F1 scores for each label within each sample. The upper part shows the label annotated for the poster section, and the lower is for the comment section.

Category	**Train**	**Validation**	**Test**
Instance_num	842	56	100
Diagnosis	5505	404	922
Treatment	1390	91	151
Problem_Current	894	56	104
Duration	535	38	62
Test	422	40	59
Age	385	24	54
Sex	350	25	54
Previous_Treatment	253	15	32
Previous_Diagnosis	122	8	9
Clinical_History	87	8	9

Table 3.7: Statistics of the dataset splits: Instance_num shows the number of threads in each dataset. The following parts are the breakdown annotations for each labels which is sorted by the numbers.

Chapter 4

EXPERIMENTS: BENCHMARKING LLMS ON MARS

In this chapter, we investigate different LLMs performances on the MARS testset. Section 4.1 shows the detailed experimental design, Section 4.2 and 4.3 give the results for each research question, and Section 4.4 provide error analysis for the previous results.

4.1 Experimental Design

We focus on benchmarking LLMs' performances on the dataset, specifically on LLM's information extraction and summarization capacities on the Chinese medical forum discussion dataset. Specifically, we want to answer two research questions based on our dataset:

(1) How well can LLMs of different sizes/types directly generate structured medical thread discussion summaries? Will zero-shot learning and dynamic few-shot learning impact the summarization performance of LLMs?

(2) Will extracting relevant information first and then generating summaries perform better than directly generating summaries with LLMs?

4.1.1 Model selection

In this experiment, we aim to assess the effectiveness of various large language models in performing information extraction and structured summarization on medical thread discussions. Specifically, we focus on evaluating each model' s ability to extract and categorize relevant information, generate structured summaries, and explore whether a sequential extraction-then-summarization approach could yield superior results compared to direct summarization.

The experiment includes five models: GPT-4o-mini [46], GLM-4-Plus [11], GLM-4-Flash [11], Qwen2-7B-Instruct [52], and Qwen2-72B-Instruct [52].

GPT-4o-mini [46] The GPT-4o-mini model from OpenAI represents a significant advancement in optimizing artificial intelligence for wider accessibility. This model is specially designed to perform effectively with reduced operational costs, priced at 15 cents per million input tokens and 60 cents per million output tokens—a substantial reduction compared to earlier models. With a context window of 128K tokens and top-notch scores on benchmarks such as the MMLU, GPT-4o-mini excels across a variety of tasks, including multimodal reasoning, which is particularly beneficial for understanding complex medical topics and generating summaries that integrate textual and visual information. Its efficiency in processing extensive input makes it a pertinent choice for summarizing medical discussions that often contain multiple threads and varied information types.

GLM-4-plus and GLM-4-flash [11] The GLM-4-plus model showcases enhanced language comprehension abilities within a multilingual context, achieving notable progress in instruction following and long text processing. These features are vital when summarizing convoluted medical discourse, where a nuanced understanding of instructions and context retention can significantly affect output quality. The model's multimodal capabilities, specifically in image and video understanding, ensure it can cater to diverse medical content, enriching the summarization process with relevant visual context. Simultaneously, GLM-4-flash is a free-access variant that democratizes access to LLM capabilities, allowing for real-time applications that can generate structured summaries without the associated costs of premium models.

Qwen2-7B-Instruct and Qwen2-72B-Instruct The Qwen2 series, particularly the Qwen2-7b-instruct, and Qwen2-72b-instruct, are remarkable for their extensive context length capabilities (up to 131K tokens) and multilingual support across 29 languages. This flexibility

enables these models to handle long, complex conversations typical in medical threads. They are designed to follow instructions effectively, making them apt for generating structured summaries from loosely connected discussions. The blend of supervised fine-tuning and preference optimization equips the Qwen2 models with improved performance across multiple language generation dimensions—key attributes when assessing how zero-shot and dynamic few-shot learning paradigms can enhance summarization efficacy.

In the landscape of large language models, **GPT-4o-mini** stands out as a benchmark for the industry. Its performance has been extensively compared across various benchmark experiments, making it one of the most widely used models. This widespread utilization underscores its reliability and effectiveness in real-world applications, particularly in tasks requiring nuanced understanding and sophisticated text generation. The **GLM-4-plus** model, recognized as the best in the GLM series, has shown a competitive edge in several tasks, demonstrating performance capabilities that can rival those of GPT-4. Furthermore, the **GLM-4-flash** model expands accessibility by providing a free-to-use option while still harnessing the advanced features and functionalities characteristic of larger models. Particular attention is paid to the **Qwen2-7B-Instruct** and **Qwen2-72B-Instruct** models, which allow for a comparative analysis of the impact of model size on performance. By investigating the differences between the 7 billion and 72 billion parameter models, this research aims to draw insightful conclusions that elucidate how scaling model size influences a model's ability to generate structured summaries effectively. This nuanced understanding will be critical in determining the optimal configurations for diverse applications, especially in the complex domain of medical discussions.

4.1.2 In-context learning strategies

In the structured summarization task, each model is tasked with generating summaries that capture the key points of medical thread discussions, such as patient symptoms, diagnoses, and treatment plans. Three learning strategies are applied to examine their effects on summarization performance. Zero-shot learning involves models generating summaries without

any examples, relying solely on the input text. Dynamic few-shot learning entails selecting relevant examples based on input content, to enhance the accuracy and completeness of the generated summaries. Specifically, we find one and three most similar threads with the test thread and give LLMs how we compose the summaries of these similar samples. Sample demonstrations are from the training and validation sets, and we only test the threads from the testing set. Similarities between samples are calculated based on the cosine similarities of threads' BERT embeddings [17]. Figure 4.1 demonstrates how these in-context learning strategies will be implemented in the experiment.

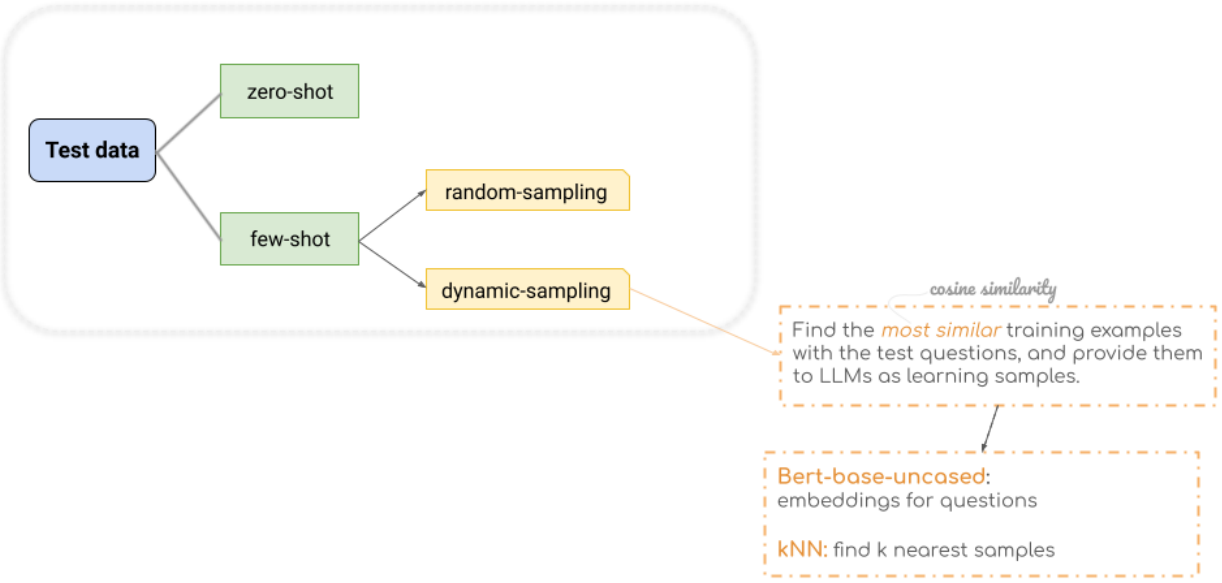

Figure 4.1: LLM In-context Learning Strategies: For few-shot learning, there are two ways to give demonstrations, random-sampling and dynamic-sampling.

To further investigate whether a two-step approach can yield better results, an additional experiment is conducted where models first extract relevant information from the medical thread discussions and then generate summaries based on this extracted content. We first ask the LLM to extract information as how we instruct annotators to extract in-

formation, and give what it extracts to ask them to generate summaries. This approach is compared to direct summarization, allowing us to determine if the preliminary extraction step contributes to generating more accurate and comprehensive summaries. And the extraction-then-summarization method also simulates the human annotation process.

While writing prompts, we refer to the framework COSTAR [16] to better utilize the power of LLMs. Appendix A shows prompts provided to the LLMs.

4.1.3 Automatic Evaluation Protocols

The evaluation protocol for this experiment is designed to rigorously measure model performance in both summarization and information extraction tasks. For summarization, ROUGE and BERTScore metrics are applied to assess the quality of the generated summaries. ROUGE measures the overlap between generated summaries and reference summaries, with a focus on recall across different n-gram levels, while BERTScore offers a more nuanced evaluation by calculating the semantic similarity between generated and reference summaries through contextual embeddings. Together, these metrics provide a robust assessment of both lexical overlap and semantic alignment.

This evaluation protocol enables a thorough examination of model performance across different tasks, learning strategies, and model sizes. By systematically analyzing summarization and information extraction outcomes, the study seeks to illuminate the factors that most significantly influence LLM effectiveness in structured medical thread discussions.

4.2 Research Question 1: Structured Summarization Performance

The experimental results presented in Table 4.1 illustrate the structured summarization performance of various large language models evaluated with different learning strategies across three metrics: ROUGE-1, ROUGE-2, and BERTScore. These performance indicators provide insight into the models' abilities to generate concise and relevant summaries from medical discussions, addressing our research question: "How well can LLMs of different sizes/types directly generate structured medical thread discussion summaries? Will zero-shot

learning and dynamic few-shot learning impact the summarization performance of LLMs?"

Model	**Learning Strategy**	**ROUGE-1**	**ROUGE-2**	**BERTScore**
GPT-4o-mini	Zero-shot	0.26	0.14	0.85
	Dynamic one-shot	0.25	0.13	0.85
	Dynamic three-shot	0.26	0.13	0.85
GLM-4-Flash	Zero-shot	0.36	0.19	0.87
	Dynamic one-shot	0.36	0.19	0.87
	Dynamic three-shot	0.36	0.18	0.87
GLM-4-Plus	Zero-shot	0.36	0.17	0.87
	Dynamic one-shot	**0.41**	**0.21**	0.88
	Dynamic three-shot	**0.41**	**0.21**	0.88
Qwen2-7B-Instruct	Zero-shot	0.25	0.06	0.83
	Dynamic one-shot	0.25	0.07	0.83
	Dynamic three-shot	0.24	0.06	0.83
Qwen2-72B-Instruct	Zero-shot	0.33	0.16	0.87
	Dynamic one-shot	0.32	0.16	0.87
	Dynamic three-shot	0.34	0.17	0.87

Table 4.1: Structured Summarization Performance Across Different Models and Learning Strategies on the MARS testset

From the results, it is evident that the GLM-4-Plus model achieved the highest scores across all learning strategies and ROUGE metrics. Specifically, it recorded a ROUGE-1 score of 0.41 and a ROUGE-2 score of 0.21 in the dynamic one-shot and three-shot settings. This indicates its strong capability to produce high-quality summaries, likely due to its advanced training and architecture. In contrast, Qwen2-7B-Instruct underperformed significantly, with the lowest ROUGE-1 and ROUGE-2 scores (0.25 and 0.06, respectively) across all learning strategies, suggesting limitations in its summarization proficiency compared to its larger

counterparts.

The GLM-4-Flash and GLM-4-Plus models presented similar performance levels, with GLM-4-Flash yielding a ROUGE-1 score of 0.36 and GLM-4-Plus achieving the same score under zero-shot conditions. Notably, both models maintained consistent performance across various learning strategies, demonstrating robustness and reliability. The BERTScore, which evaluates semantic similarity, corroborated these findings, with the best BERTScore of 0.88 from GLM-4-Plus using dynamic one-shot or three-shot learning strategies, showcasing its superior ability to generate coherent and contextually relevant summaries.

The findings suggest that larger models generally outperform smaller ones in generating structured summaries, supporting the hypothesis that model size and architecture significantly affect performance. The GLM-4-Plus's improved scores with dynamic learning strategies reflect its advantage in adaptability, suggesting that utilizing few-shot learning can enhance the summarization process, particularly in complex domains like medical discussions.

In contrast, the GPT-4o-mini model exhibited consistent but lower performance compared to the GLM-4 variants, especially in zero-shot conditions where it scored 0.26 (ROUGE-1). While the GPT-4o-mini is a cost-effective option, its inability to surpass the results of the GLM-4 series highlights a potential trade-off between accessibility and performance.

Overall, this experiment indicates that GLM-4-Plus emerges as a potential industry leader in structured summarization tasks within the medical discourse context. The significant performance gap between the models also demonstrates that employing dynamic few-shot learning strategies can positively influence summarization effectiveness, particularly for larger models.

4.3 Research Question 2: Comparison of Direct Summarization vs. Extraction-then-Summarization

The results presented in Table 4.2 focus on the extraction-then-summarization approach across various large language models and learning strategies. The performance is assessed

using three metrics: ROUGE-1, ROUGE-2, and BERTScore. This analysis addresses the second research question: "Will extracting relevant information first and then generating summaries perform better than directly generating summaries with LLMs?"

Table 4.2: Extraction-then-Summarization Performance Across Different Models and Learning Strategies on the MARS testset

Model	**Learning Strategy**	**ROUGE-1**	**ROUGE-2**	**BERTScore**
GPT-4o-mini	Zero-shot	0.21	0.10	0.87
	Dynamic one-shot	0.37	0.17	0.90
GLM-4-Flash	Zero-shot	0.19	0.08	0.86
	Dynamic one-shot	0.31	0.14	0.89
GLM-4-Plus	Zero-shot	0.30	0.12	0.89
	Dynamic one-shot	**0.44**	0.21	0.92
Qwen2-7B-Instruct	Zero-shot	0.05	0.02	0.82
	Dynamic one-shot	0.25	0.09	0.86
Qwen2-72B-Instruct	Zero-shot	0.19	0.09	0.87
	Dynamic one-shot	0.40	0.18	0.91

The findings indicate that utilizing an extraction-then-summarization methodology generally yields improved performance compared to a direct summarization approach, particularly for most models when using dynamic one-shot learning. For instance, GPT-4o-mini reaches a ROUGE-1 score of 0.37 with dynamic one-shot learning compared to its zero-shot score of 0.21. Similarly, GLM-4-Plus achieves a remarkable ROUGE-1 score of 0.44 with dynamic one-shot learning, which is significantly higher than its zero-shot score of 0.30.

The improved performance across models suggests that extracting relevant information prior to summary generation can allow the models to focus on the most pertinent details, leading to higher-quality outputs. The BERTScore metrics further support this trend, with GLM-4-Plus achieving the highest score of 0.92 when adopting the extraction-then-

summarization approach. This indicates that the generated summaries not only capture the required content better but also maintain semantic relevance to the source material.

When comparing these results with the direct summarization performance from the first table (as shown in Figure 4.2), it is evident that the extraction-then-summarization approach generally outperforms the direct summarization method. For example, GLM-4-Plus showed substantial improvement, achieving a ROUGE-1 score of 0.44 (dynamic one-shot) compared to the direct zero-shot score of 0.36. The GPT-4o-mini had a ROUGE-1 score of 0.37 in the extraction scheme versus 0.26 in direct summarization.

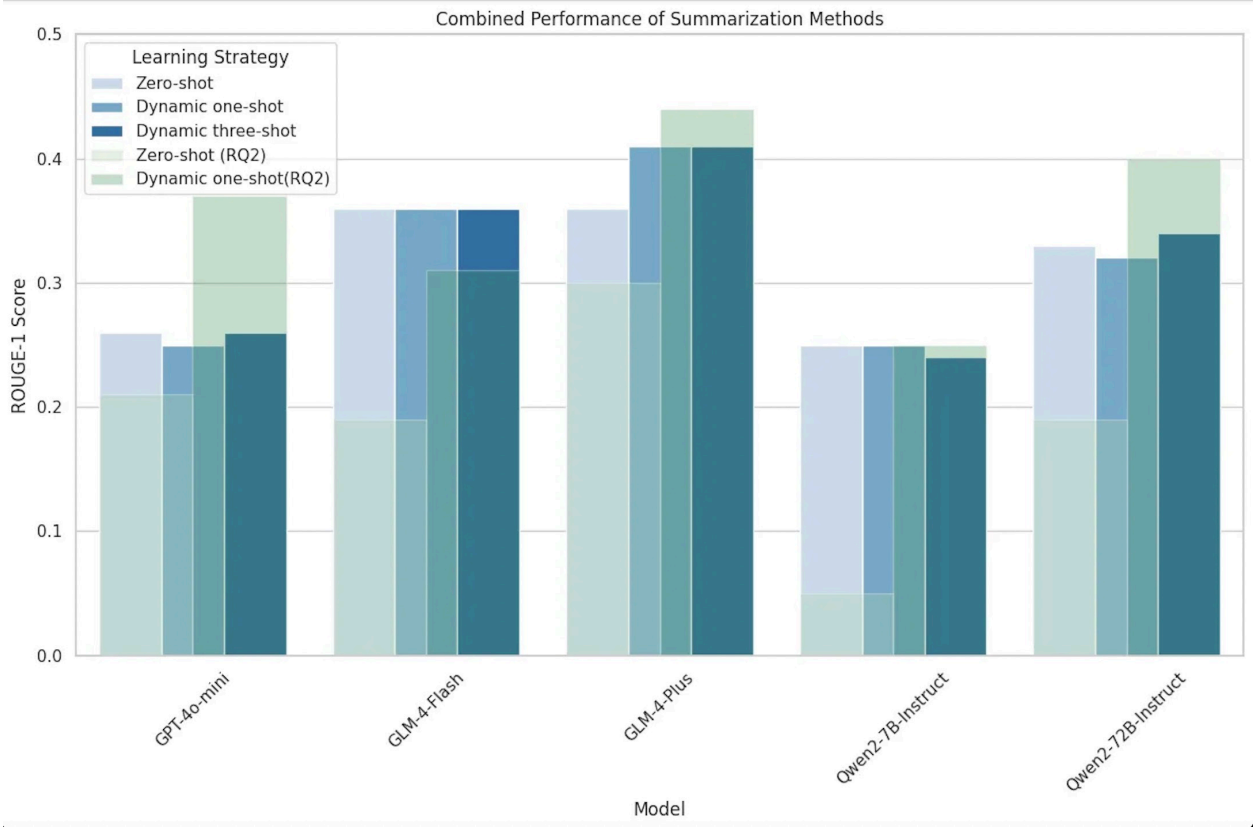

Figure 4.2: Visualization of summarization performances of two research questions.

Conversely, the Qwen2-7B-Instruct model illustrates the challenges of a less effective information extraction task. With a zero-shot ROUGE-1 score of only 0.05 in the extraction

scenario and no notable improvement in its dynamic one-shot strategy, it indicates that poor performance in the extraction phase leads to subpar summary generation. This reinforces the notion that if the initial extraction task fails to identify relevant information adequately, the subsequent summarization process will likely suffer as well.

In conclusion, the results suggest that extracting relevant information first and then generating summaries is generally more effective than directly generating summaries with LLMs, especially for models with robust capabilities. This analysis highlights that the extraction phase serves as a critical component: when executed well, it provides a solid foundation for effective summary generation. Conversely, if the extraction process falls short, as seen with the Qwen2-7B-Instruct model, the benefits of summarization may be significantly diminished. Therefore, ensuring a high-quality information extraction step is essential for enhancing the overall performance of summarization tasks in complex domains like medical discussions.

4.4 Error Analysis

From the experimental results, several questions are worth speculating:

(1) Why is there a huge gap between the BERTScore and the ROUGE score?

(2) Why for some cases, in-context learning has adverse impacts on the results?

(3) Why is the impact of extraction-then-summarization polarized, with negative effects in zero-shot scenarios and positive effects in one-shot scenarios, and why does it seem to be more beneficial for larger or better-performing models?

4.4.1 Error Analysis 1

The gap between BERTScore and ROUGE scores stems from their fundamentally different measurement approaches, which often results in significant discrepancies in evaluation outcomes. ROUGE primarily focuses on n-gram overlap, assessing how many words or phrases

present in the reference summaries are also found in the generated summaries. This method can overlook semantic nuances, as it is highly dependent on exact word matches. Figure 4.3 shows an example that has a low ROUGE score but a high BERTS score.

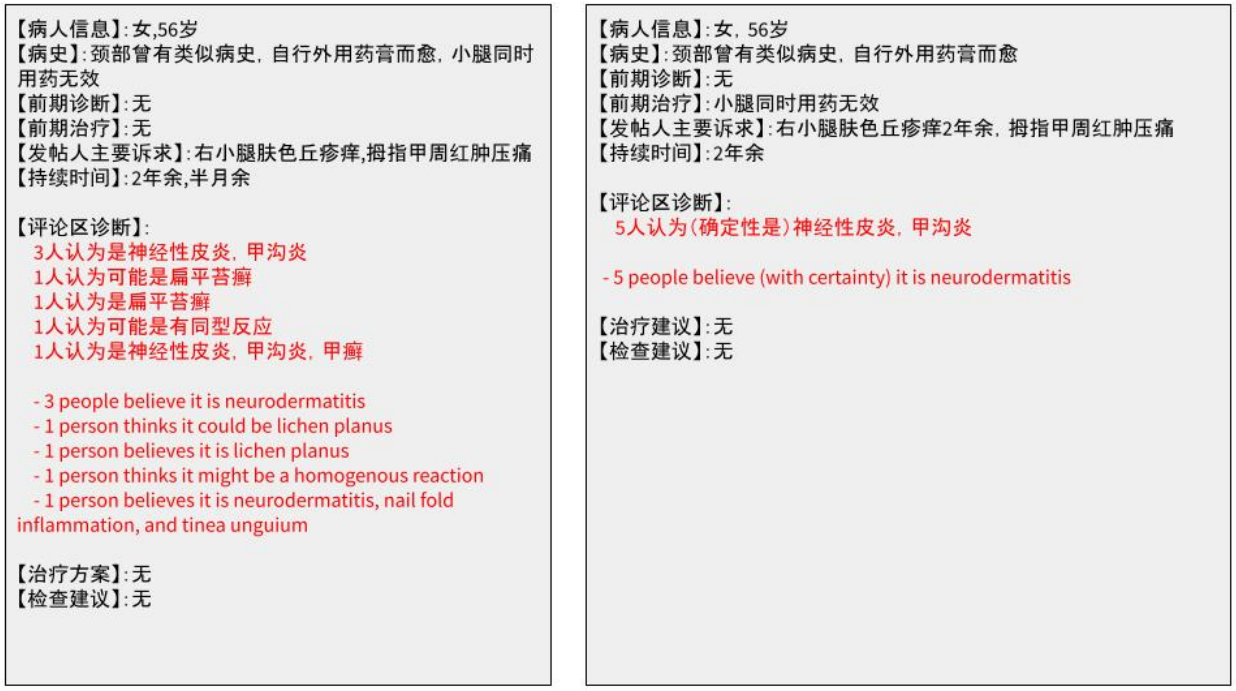

Figure 4.3: An example that has a low ROUGE score but a high BERTS score: the left side is the gold summary and the right side is the system summary. The part in red is the main reason for the gap.

While ROUGE may score these summaries low due to minimal n-gram overlap, BERTScore would recognize the semantic equivalence between terms like ”neurodermatitis” and ”skin issues,” thus yielding a higher score. This clearly demonstrates how varying perspectives on text can lead to different scoring outcomes and highlights the importance of using diverse metrics for evaluating summarization quality.

In the provided analysis and example, it is evident that while both metrics serve as useful tools for evaluating the quality of summaries, their differing methodologies can produce conflicting results. Understanding these discrepancies is valuable for researchers and

practitioners when interpreting the effectiveness of summarization algorithms.

Additionally, we also find that the summary we constructed is structured and contains many fixed entries (such as Patient Information, Medical History, etc.). In fact, for some data, the content of these fixed-format entries can even exceed the length of the summary itself. Based on the prompt and examples, LLMs are very adept at producing summaries with similar structures. However, the differences between the generated summaries and the gold summary are minor yet valuable , such as the statistical tally of the number of people for certain diagnoses, as well as the judgment of **Diagnosis_Val** (whether it is present, possible, negated, or conditional). We also observed that large models tend to label diagnoses as present or negated, but sometimes the uncertain tone of the commentators, or even implications, is difficult to capture. On the other hand, for categories such as **Patient Information** that are specific, it is easy for LLMs to capture them right. Hence, for evaluation metrics, the semantics that are presented by embeddings might be quite similar between the gold summaries and the generated ones. Therefore, BERTScore can provide a rough evaluation, especially when the performance differences between models are quite pronounced. However, when models are able to effectively follow the prompt to generate seemingly correct or compliant summaries, using ROUGE can better reveal the performance differences between the models.

4.4.2 Error Analysis 2

In-context learning strategies are aimed to aid LLMs' performances. However, in our experimental results, specifically the ROUGE-1 score, in some cases in-context learning posts inverse results on LLM performances. We need to acknowledge that although there are instances where in-context learning causes a decrease in ROUGE-1 scores, such occurrences are relatively infrequent and the differences are not significant. In the Structured Summarization task, the ROUGE-1 score for GPT-4o-mini in zero-shot is 0.26, while in dynamic one-shot it is 0.25. For Qwen2-7B-Instruct, the ROUGE-1 score in dynamic one-shot is 0.25, and in dynamic three-shot it is 0.24 (as shown in Table 4.1. There may be two possible

situations that lead to the decrease: (1) corner cases (as exemplified in the following 4.4.2 with over-complicated discussion contents);

Gold Summary

病人信息】: 男,16岁

【病史】: 脚上的皮损是烫

【前期诊断】: 慢性湿疹,也有怀疑疥疮的

【前期治疗】: 按疥来治了一次,无效,曾用抗菌素治疗会好转

【发帖人主要诉求】: 手足脓疱,头上曾起过,现在主要是掌跖及手足侧缘,再一看这病还是湿疹吗,这些小脓疱怎么解释?

【持续时间】: 10年

【评论区诊断】:

2人认为是湿疹感染

1人认为可能是慢性湿疹合并感染

1人认为可能是手上的像是湿疹, 脓疱疮, 脚上的像剥脱性角质松懈症

1人认为可能是掌跖脓疱病

1人认为可能是白塞病

1人认为可能是真菌感染

1人认为可能是神经性皮炎

1人认为是脓疱疮

1人认为是足癣并癣菌疹

1人认为是连续性指端皮炎

【治疗方案】:

无

【检查建议】:

真菌检查 (基于诊断: 是足癣并癣菌疹)

GPT-4o-mini (zero-shot)

【病人信息】: 16岁, 男

【病史】: 无

【前期诊断】: 慢性湿疹, 曾怀疑疥疮

【前期治疗】: 曾按疥疮治疗无效, 曾用抗菌素治疗会好转

【发帖人主要诉求】: 询问手足脓疱的原因及是否仍为湿疹

【持续时间】: 10年

【评论区诊断】:

7人认为可能是“掌跖脓疱病”

1人认为是脓疱疮

1人认为是湿疹感染

1人认为是足癣并癣菌疹

1人认为是连续性指端皮炎

1人认为是神经性皮炎

【治疗建议】:

无

【检查建议】:

建议做真菌检查 (基于诊断: 足癣并癣菌疹)

GPT-4o-mini (dynamic one-shot)

【病人信息】: 男, 16岁

【病史】: 无

【前期诊断】: 慢性湿疹, 曾怀疑疥疮

【前期治疗】: 曾按疥疮治疗无效, 使用抗菌素治疗有好转

【发帖人主要诉求】: 询问手足脓疱的原因及是否仍为湿疹

【持续时间】: 10年

【评论区诊断】：8人认为可能是掌跖脓疱病
【治疗建议】：无
【检查建议】：建议做真菌检查（基于诊断：足癣并癣菌疹）

(2) For smaller LLMs like 7B, an excess of prompts might prevent them from fully adhering to the instructions effectively (as examplified in the following 4.4.2.)

Gold Summary
【病人信息】：无,10月大
【病史】：无
【前期诊断】：无
【前期治疗】：无
【发帖人主要诉求】：左下腹部白色扁平丘疹、斑块,无病理，请教,诊断
【持续时间】：7个月

【评论区诊断】：
3人认为可能是疣状痣
2人认为是疣状痣
1人认为可能是线状表皮痣
【治疗方案】：
无
【检查建议】：
无

Qwen2-7B-Instruct (dynamic one-shot)
【病人信息】：年龄：10月大；性别：未提及
【病史】：无
【前期诊断】：无

【前期治疗】：无
【发帖人主要诉求】：询问或讨论左下腹部白色扁平丘疹、斑块的诊断
【持续时间】：7个月

【评论区诊断】：
5人认为是疣状痣
【治疗建议】：
无具体治疗建议
【检查建议】：
无具体检查建议

Qwen2-7B-Instruct (dynamic three-shot)
【病人信息】：无年龄、性别信息
【病史】：无其他疾病信息
【前期诊断】：线状表皮痣、疣状痣
【前期治疗】：无治疗信息
【发帖人主要诉求】：询问或讨论左下腹部白色扁平丘疹、斑块的诊断
【持续时间】：7个月

【评论区诊断】：
5人认为是疣状痣
【治疗建议】：
无具体治疗建议
【检查建议】：
无具体检查建议

4.4.3 Error Analysis 3

The performances of large language models are closely related to their extraction capabilities. Generally, larger LLMs (GLM-4-Plus in our experiments) exhibit better performance when responding to detailed and complex extraction prompts (as shown in Table 4.2), which differ significantly from traditional information extraction tasks, such as entity and relation extraction. On the other hand, small LLMs (Qwen2-7B-Instruct in our experiments) might not be good at unfamiliar and complicated tasks.

In the case of larger LLMs, extraction tasks serve as an aid to their performance. Conversely, for smaller LLMs, these extraction tasks might act as a distraction, potentially hindering their overall effectiveness.

4.5 Conclusions for Research Questions

RQ1: How well can LLMs of different sizes/types directly generate structured medical thread discussion summaries? Will zero-shot learning and dynamic few-shot learning impact the summarization performance of LLMs? The evaluation of LLMs of varying sizes and types in generating structured medical thread discussion summaries reveals a strong correlation between model size and summarization performance. In our experiments, the GLM-4-Plus model consistently achieved the highest ROUGE and BERTScore metrics across all learning strategies, particularly excelling in dynamic one-shot and three-shot settings. This model's architecture likely contributes to its ability to produce high-quality summaries that capture the intricacies of medical discussions. In contrast, smaller models like Qwen2-7B-Instruct demonstrated significantly lower performance, indicating their limitations in summarization proficiency. The findings suggest that larger models, especially when utilizing dynamic few-shot learning strategies, are better suited for generating coherent and contextually relevant summaries. Overall, zero-shot learning and dynamic few-shot learning strategies generally enhance the summarization capabilities of LLMs, with the latter showing notable improvements in model adaptability and perfor-

mance.

RQ2: Will extracting relevant information first and then generating summaries perform better than directly generating summaries with LLMs? The analysis comparing extraction-then-summarization approaches to direct summarization indicates that the former generally yields superior results. Extraction-then-summarization not only enhances the quality of the generated summaries but also allows the models to focus on pertinent details, especially in terms of few-shot cases, leading to improved performance metrics across most evaluated models. Specifically, in the dynamic one-shot configuration, notable improvements were observed, with the GLM-4-Plus model achieving a ROUGE-1 score of 0.44 under extraction-then-summarization compared to a direct score of 0.36. This evidence underscores the notion that effectively extracting relevant information prior to summarization significantly aids in capturing the essence of medical discussions. Conversely, smaller models often struggled in the extraction phase, demonstrating a clear limitation in their overall performance. Thus, the results emphasize the critical importance of a robust extraction step in the summarization process, especially for complex and nuanced medical dialogues.

Chapter 5

CONCLUSION

In the discussion chapter, we summarize our main contributions and future work. In the last section, we reach the final conclusion of the study.

5.1 Main Contributions

In this work, we introduce the first dataset specifically designed for summarizing medical forum discussions, referred to as the MARS (MedicAl thRead Summarization) dataset. The main contributions of this research are as follows:

Dataset Contributions The **MARS** dataset is carefully designed to address the complexities of medical dialogue from multiple dimensions, supporting both **information extraction** and **summarization tasks**:

- **Cleaned Forum Content:** It builds on the cleaned forum data shared with DermaVQA [67], providing a solid foundation for subsequent analyses (see Figure 1.1).
- **Advanced Annotations:** The dataset introduces nuanced annotations that go beyond conventional entity and relation extraction methods. These annotations identify complex phrases related to treatment and diagnosis, presenting a more detailed view of medical discourse (as shown in Table 1.1).
- **Structured Summaries:** Summaries derived from expert annotations encapsulate essential elements of medical discussions, such as patient conditions, diagnoses, treatment plans, and recommendations (see Figure 1.2). These summaries provide clear and organized insights into the discussions.

By thoughtfully crafting the MARS dataset, we aim to facilitate advances in **information extraction** and **summarization** within the medical domain, ultimately supporting the development of more effective healthcare applications and research methodologies.

Benchmark Contributions The MARS dataset also serves as a valuable **benchmark** for evaluating large language models in the medical domain:

- The tasks in the MARS dataset introduce domain specific challenges, requiring LLMs to engage in advanced reasoning and interpret implications in complex medical discussions.

- The dataset provides a comprehensive evaluation framework for both IE and summarization, addressing the dual goals of extracting relevant information and generating coherent, insightful summaries.

- As a resource specifically designed for medical discourse, the MARS dataset fills a critical gap in the existing literature, paving the way for future advancements in automated summarization and enhancing LLM performance in real-world healthcare scenarios.

Through these contributions, the MARS dataset promotes robust research in medical discourse, fostering the development of LLMs tailored to the intricacies of the medical domain and ultimately supporting better healthcare applications.

5.2 Future works

While the MARS dataset represents a significant step forward in the field of medical discourse summarization, several limitations should be acknowledged. One major limitation is the current evaluation metrics employed to assess the quality of generated summaries. Metrics such as ROUGE and BERTScore, although commonly used, may not fully capture the semantic richness and specific contextual requirements inherent in medical discussions. For

example, these metrics often rely on n-gram matching, which can overlook nuanced differences in terminology that are valuable in a healthcare context. Thus, there is a pressing need for the development of better evaluation metrics that can appropriately reflect the quality and reliability of summaries generated from medical dialogue.

Looking to the future, there are numerous avenues for enhancing the MARS dataset and the methodologies used in summarization tasks. One promising direction is the exploration of supervised fine-tuning (SFT) of LLMs on our dataset. By training these models specifically on our medical forum content, we can potentially improve their ability to perform information extraction and summarization tasks effectively. Fine-tuning could also help in overcoming some of the limitations observed in our current experiments, leading to more coherent and medically relevant outputs.

Furthermore, we are in the process of developing an English version of the MARS dataset, which will expand its accessibility and usability for a global audience of researchers and practitioners. This English version aims to retain the rich details and context of the original dataset while making it easier for non-Chinese speakers to engage with our research and findings. By addressing these limitations and expanding the dataset, we hope to catalyze further research and development in the area of automated medical discourse summarization.

5.3 Conclusion

In conclusion, the introduction of the MARS dataset establishes a foundational resource for improving the automated summarization of medical forum discussions, addressing the unique needs and complexities of healthcare dialogues that have been largely overlooked in prior research. The evaluation of large language models on the MARS dataset reveals that larger models, particularly the GLM-4-Plus, significantly outperform smaller counterparts in generating structured summaries, especially when utilizing dynamic few-shot learning strategies. Furthermore, our findings indicate that an extraction-then-summarization approach yields better performance compared to direct summarization methods, underscoring the importance of effective information extraction as a precursor to high-quality summarization.

Despite the significant strides made, including the promotion of research tailored to medical discourse and the establishment of a benchmark for evaluating LLM performance, our study acknowledges its limitations, particularly regarding the adequacy of current evaluation metrics. The implications of our findings underscore the necessity for more sophisticated metrics capable of effectively evaluating the nuances of medical summaries. Moreover, we advocate for future research directions, such as the supervised fine-tuning of LLMs on the MARS dataset and the development of an English version to broaden its accessibility. Overall, our work serves as a catalyst for further advancements in automated medical discourse summarization, paving the way for improved healthcare communication and access to critical patient information.

BIBLIOGRAPHY

[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors, 2022.

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[3] Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li, editors, *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 239–251, Singapore and Online, July 2021. Association for Computational Linguistics.

[4] Emma Barker and Robert Gaizauskas. Summarizing multi-party argumentative conversations in reader comment on news. In Chris Reed, editor, *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 12–20, Berlin, Germany, August 2016. Association for Computational Linguistics.

[5] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[6] Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online, June 2021. Association for Computational Linguistics.

[7] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the*

Association for Computational Linguistics, pages 2291–2302, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[8] Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. An investigation of evaluation methods in automatic medical note generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. An investigation of evaluation methods in automatic medical note generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada, July 2023. Association for Computational Linguistics.

[10] Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. Summarizing online forum discussions – can dialog acts of individual messages help? In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2131, Doha, Qatar, October 2014. Association for Computational Linguistics.

[11] BigModel. How to use the model. Accessed: 2024-11-18.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[14] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[15] Cohere. Introduction to large language models. Accessed: 2024-11-11.

[16] DataStax. Ragstack default architecture. Accessed: 2024-11-05.

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

[18] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: General language model pretraining with autoregressive blank infilling. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[19] Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In Parminder Bhatia, Steven Lin, Rashmi Gangadharaiah, Byron Wallace, Izhak Shafran, Chaitanya Shivade, Nan Du, and Mona Diab, editors, *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online, July 2020. Association for Computational Linguistics.

[20] Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), pages 6866–6880, Online, August 2021. Association for Computational Linguistics.

[21] Alexander Fabbri, Xiaojian Wu, Srini Iyer, Haoran Li, and Mona Diab. AnswerSumm: A manually-curated dataset and pipeline for answer summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2508–2520, Seattle, United States, July 2022. Association for Computational Linguistics.

[22] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution, 2023.

[23] Marcelo Fiszman, ThomasC. Rindflesch, and Halil Kilicoglu. Summarizing drug information in medline citations. Dec 2005.

[24] Yang Gao, Wei Zhao, and Steffen Eger. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online, July 2020. Association for Computational Linguistics.

[25] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics.

[26] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. LongT5: Efficient text-to-text transformer for long sequences. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States, July 2022. Association for Computational Linguistics.

[27] Bernal Jiménez Gutiérrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about gpt-3 in-context learning for biomedical ie? think again, 2022.

[28] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend.

In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA, 2015. MIT Press.

[29] Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. A survey on multi-modal summarization, 2023.

[30] Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online, November 2020. Association for Computational Linguistics.

[31] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics, and Natural Language Processing.* Prentice Hall, Saddle River, NJ, 2008.

[32] Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online, August 2021. Association for Computational Linguistics.

[33] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics.

[34] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[35] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Meeting of the Association for Computational Linguistics,Meeting of the Association for Computational Linguistics*, Jul 2004.

[37] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. BRIO: Bringing order to abstractive summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[38] Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. Multi-document summarization via deep learning techniques: A survey, 2021.

[39] Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37, 2016.

[40] Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. Gpt-3 models are poor few-shot learners in the biomedical domain, 2022.

[41] Milad Moradi and Nasser Ghadiri. Different approaches for identifying important concepts in probabilistic biomedical text summarization. *Artificial Intelligence in Medicine,Artificial Intelligence in Medicine*, May 2016.

[42] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: a recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press, 2017.

[43] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Stefan Riezler and Yoav Goldberg, editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics.

[44] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[45] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023.

[46] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-11-18.

[47] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres,

Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

[48] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[49] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[50] Keighley Overbay, Jaewoo Ahn, Fatemeh Pesaran zadeh, Joonsuk Park, and Gunhee Kim. mRedditSum: A multimodal abstractive summarization dataset of Reddit threads with images. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4117–4132, Singapore, December 2023. Association for Computational Linguistics.

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method

for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[52] Qwen2. Qwen2. Accessed: 2024-11-18.

[53] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020.

[54] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[55] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics.

[56] Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto, and Juneyoung Park. Dialogue summaries as dialogue states (DS2), template-guided summarization for few-shot dialogue state tracking. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3824–3846, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[57] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[58] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

[59] Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. Summarizing medical conversations via identifying important utterances. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*,

pages 717–729, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[60] Xiangru Tang, Arman Cohan, and Mark Gerstein. Aligning factual consistency for clinical studies summarization through reinforcement learning. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 48–58, Toronto, Canada, July 2023. Association for Computational Linguistics.

[61] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[63] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470, 2020.

[64] Wen Xiao and Giuseppe Carenini. Extractive summarization of long documents by combining global and local context. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3011–3021, Hong Kong, China, November 2019. Association for Computational Linguistics.

[65] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers, 2023.

[66] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586, Sep 2023.

[67] Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. DermaVQA: A Multilingual Visual Question Answering Dataset for Dermatology . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland, October 2024.

[68] Ww. Yim, Y. Fu, A. Ben Abacha, and et al. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Sci Data*, 10:586, 2023.

[69] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation, 2021.

[70] Amy Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing online discussion using coarse discourse sequences. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):357–366, May 2017.

[71] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR, 13–18 Jul 2020.

[72] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.

[73] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022.

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

[75] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024.

[76] Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. MIE: A medical information extractor towards medical dialogues. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469, Online, July 2020. Association for Computational Linguistics.

[77] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In Dan Jurafsky, Joyce Chai,

Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online, July 2020. Association for Computational Linguistics.

Appendix A
LLM PROMPTING

A.1 Zero-shot direct summarization

#CONTEXT#
我希望你帮我总结以下医疗论坛讨论的帖子，抽取相关信息，并生成结构性的摘要。
我将给你提供<post_title>, <post_content>, <comments>,
提供内容中也包含发帖人和评论人的id，有时发帖人也会在评论区回复评论或补充信息。
我希望你从以下几个方面在生成摘要。
-【病人信息】：包括病人的年龄、性别
-【病史】：病人曾经的疾病（不包含当前在帖子里询问的特定疾病）
-【前期诊断】：对当前询问疾病的曾经的诊断
-【前期治疗】：对当前询问疾病的曾经的治疗，若有前期治疗效果也需要包含
-【发帖人主要诉求】：本帖主要想询问或讨论的内容（包含对当前疾病情况的描述）
-【持续时间】当前询问的疾病的持续时间
-【评论区诊断】评论区对于发帖人询问的诊断，某些发帖人可能会给出相同的诊断，
并且他们的诊断有确定性的区分（是、可能是、不是、基于...条件是），
所以给出评论区诊断时你应该整理信息，
形成"N人认为(确定性xxx)(诊断内容xxx)"这样结构的摘要
-【治疗方案】：一些评论人可能会给出进一步治疗的建议，如果该治疗建议时基于某诊断，
请以"治疗建议（基于诊断：xxx）"的形式总结
-【检查建议】：一些评论人可能会给出进一步检查的建议，如果该检查建议时基于某诊断，
请以”检查建议（基于诊断：xxx）"的形式总结

以上【】中的条目不一定在每个帖子中全部存在，在生成摘要时若某条目不存在，请以“【条目】：无”的方式写出，不要省略这些条目。

#OBJECTIVE#

现在我将给出论坛帖子，请仔细阅读并总结。

{content}

#STYLE#

你应该尽可能直接从帖子中抽取相关信息生成摘要，最好不要转写。

#TONE#

客观抽取信息，不要给出评论。

#AUDIENCE#

这个摘要可以帮助人们更好去理解论坛帖子，使信息更整洁且有结构。

#RESPONSE#

你需要按照以下格式生成摘要:

【病人信息】：xxx

【病史】：xxx

【前期诊断】：xxx

【前期治疗】：xxx

【发帖人主要诉求】：xxx

【持续时间】：xxx

【评论区诊断】：

xx人xxx

【治疗建议】：

xxx
【检查建议】：
xxx

注意，只需要输出以上摘要，不需要给出任何其他的解释或者评论或任何你思考的过程。

A.2 Few-shot direct summarization

#CONTEXT#
我希望你帮我总结以下医疗论坛讨论的帖子，抽取相关信息，并生成结构性的摘要。
我将给你提供<post_title>, <post_content>, <comments>,
提供内容中也包含发帖人和评论人的id，有时发帖人也会在评论区回复评论或补充信息。
我希望你从以下几个方面在生成摘要。
-【病人信息】：包括病人的年龄、性别
-【病史】：病人曾经的疾病（不包含当前在帖子里询问的特定疾病）
-【前期诊断】：对当前询问疾病的曾经的诊断
-【前期治疗】：对当前询问疾病的曾经的治疗，若有前期治疗效果也需要包含
-【发帖人主要诉求】：本帖主要想询问或讨论的内容（包含对当前疾病情况的描述）
-【持续时间】当前询问的疾病的持续时间
-【评论区诊断】评论区对于发帖人询问的诊断，某些发帖人可能会给出相同的诊断，
并且他们的诊断有确定性的区分（是、可能是、不是、基于...条件是），
所以给出评论区诊断时你应该整理信息，
形成"N人认为(确定性xxx)(诊断内容xxx)"这样结构的摘要
-【治疗方案】：一些评论人可能会给出进一步治疗的建议，如果该治疗建议时基于某诊断，
请以"治疗建议（基于诊断：xxx）"的形式总结
-【检查建议】：一些评论人可能会给出进一步检查的建议，如果该检查建议时基于某诊断，
请以”检查建议（基于诊断：xxx）"的形式总结

以上【】中的条目不一定在每个帖子中全部存在，在生成摘要时若某条目不存在，

请以“【条目】：无”的方式写出，不要省略这些条目。

你可以参照以下我提供给你的例子:
{example}

#OBJECTIVE#
现在我将给出论坛帖子，请仔细阅读并总结。
{content}

#STYLE#
你应该尽可能直接从帖子中抽取相关信息生成摘要，最好不要转写。

#TONE#
客观抽取信息，不要给出评论。

#AUDIENCE#
这个摘要可以帮助人们更好去理解论坛帖子，使信息更整洁且有结构。

#RESPONSE#
你需要按照以下格式生成摘要:
【病人信息】：xxx
【病史】：xxx
【前期诊断】：xxx
【前期治疗】：xxx
【发帖人主要诉求】：xxx
【持续时间】：xxx

【评论区诊断】：

xx人xxx

【治疗建议】：

xxx

【检查建议】：

xxx

注意，只需要输出以上摘要，不需要给出任何其他的解释或者评论或任何你思考的过程。

A.3 Zero-shot information extraction

#CONTEXT#

我希望你帮我总结以下医疗论坛讨论的帖子，

抽取相关信息我将给你提供<post_title>, <post_content>, <comments>,

提供内容中也包含发帖人和评论人的id，有时发帖人也会在评论区回复评论或补充信息。

对于发帖人和评论人需要抽取的信息是不同的。

发帖人需要提取的信息有：

- Problem_Current：本帖主要想询问或讨论的内容（包含对当前疾病情况的描述）
- Age：病人年龄，不一定局限于年龄的直接表达，“青年、老年”等也适用
- Sex：病人性别，不一定局限于性别的直接表达，“女儿、弟弟”等也适用
- Clinical_History：病人曾经的疾病（不包含当前在帖子里询问的特定疾病）
- Previous_Diagnosis：对当前询问疾病的曾经的诊断
- Previous_Treatment：对当前询问疾病的曾经的治疗
- Duration：当前询问的疾病的持续时间

评论人需要提取的信息有：

- Diagnosis：评论人提出的诊断意见，

对于诊断意见在抽取时我们还应该判断发帖人对自己诊断的确定性，

从（present, negated, possible, conditional）中选取一个他们的确定性判断的值

- Treatment: 一些评论人可能会给出进一步治疗的建议，
如果该治疗建议时基于某诊断也需要提取出
- Test: 一些评论人可能会给出进一步检查的建议，如果该检查建议时基于某诊断也需要提取出

不一定每一类信息都会出现在帖子中，另外同一个类别的内容不一定是连续出现的，
你需要变阅读内容边做抽取判断。

#OBJECTIVE#
现在我将给出论坛帖子，请仔细阅读并总结。
{content}

#STYLE#
你应该尽可能直接从帖子中抽取相关信息，最好不要转写。

#TONE#
客观抽取信息，不要给出评论。

#AUDIENCE#
抽取的信息可以帮助人们更好去理解论坛帖子，使信息更整洁且有结构。

#RESPONSE#
你需要按照以下格式输出你所抽取的信息:
类别名: 抽取的相关内容

对于发帖人的信息抽取来说，每个类别名为1行，你需要将同类别抽取出的内容拼接到一起。

对于评论人的信息抽取来说，每个评论人为一个单位进行抽取，参照以下方式输出:
评论人id

类别名：抽取的相关内容（Test/Treatment基于的诊断）

不存在的类别可以不用输出。
注意，只需要输出抽取的信息，不需要给出任何其他的解释或者评论或任何你思考的过程。

A.4 Few-shot information extraction

#CONTEXT#
我希望你帮我总结以下医疗论坛讨论的帖子，
抽取相关信息我将给你提供<post_title>, <post_content>, <comments>,
提供内容中也包含发帖人和评论人的id，有时发帖人也会在评论区回复评论或补充信息。
对于发帖人和评论人需要抽取的信息是不同的。

发帖人需要提取的信息有：
- Problem_Current：本帖主要想询问或讨论的内容（包含对当前疾病情况的描述）
- Age：病人年龄，不一定局限于年龄的直接表达，“青年、老年”等也适用
- Sex：病人性别，不一定局限于性别的直接表达，“女儿、弟弟”等也适用
- Clinical_History：病人曾经的疾病（不包含当前在帖子里询问的特定疾病）
- Previous_Diagnosis：对当前询问疾病的曾经的诊断
- Previous_Treatment：对当前询问疾病的曾经的治疗
- Duration：当前询问的疾病的持续时间

评论人需要提取的信息有：
- Diagnosis：评论人提出的诊断意见，
对于诊断意见在抽取时我们还应该判断发帖人对自己诊断的确定性，
从（present, negated, possible, conditional）中选取一个他们的确定性判断的值
- Treatment：一些评论人可能会给出进一步治疗的建议，
如果该治疗建议时基于某诊断也需要提取出
- Test：一些评论人可能会给出进一步检查的建议，如果该检查建议时基于某诊断也需要提取出

不一定每一类信息都会出现在帖子中，另外同一个类别的内容不一定是连续出现的，
你需要变阅读内容边做抽取判断。

你可以参照以下我提供给你的例子学习如何抽取，
但最后呈现出来的格式你需要按照#RESPONSE#中的规定：
{ie_example}

#OBJECTIVE#
现在我将给出论坛帖子，请仔细阅读并总结。
{content}

#STYLE#
你应该尽可能直接从帖子中抽取相关信息，最好不要转写。

#TONE#
客观抽取信息，不要给出评论。

#AUDIENCE#
抽取的信息可以帮助人们更好去理解论坛帖子，使信息更整洁且有结构。

#RESPONSE#
你需要按照以下格式输出你所抽取的信息：
类别名：抽取的相关内容

对于发帖人的信息抽取来说，每个类别名为1行，你需要将同类别抽取出的内容拼接到一起。

对于评论人的信息抽取来说，每个评论人为一个单位进行抽取，参照以下方式输出：

评论人id
类别名: 抽取的相关内容 (Test/Treatment基于的诊断)

不存在的类别可以不用输出。
注意, 只需要输出抽取的信息, 不需要给出任何其他的解释或者评论或任何你思考的过程。

A.5 Zero-shot extraction-then-summarization

#CONTEXT#
我希望你帮我总结以下医疗论坛讨论的帖子, 抽取相关信息, 并生成结构性的摘要。
我将给你提供<post_title>, <post_content>, <comments>,
提供内容中也包含发帖人和评论人的id, 有时发帖人也会在评论区回复评论或补充信息。
我希望你从以下几个方面在生成摘要。
-【病人信息】: 包括病人的年龄、性别
-【病史】: 病人曾经的疾病 (不包含当前在帖子里询问的特定疾病)
-【前期诊断】: 对当前询问疾病的曾经的诊断
-【前期治疗】: 对当前询问疾病的曾经的治疗, 若有前期治疗效果也需要包含
-【发帖人主要诉求】: 本帖主要想询问或讨论的内容 (包含对当前疾病情况的描述)
-【持续时间】当前询问的疾病的持续时间
-【评论区诊断】评论区对于发帖人询问的诊断, 某些发帖人可能会给出相同的诊断,
并且他们的诊断有确定性的区分 (是、可能是、不是、基于...条件是),
所以给出评论区诊断时你应该整理信息, 形成"N人认为(确定性xxx)(诊断内容xxx)"这样结构的摘要
-【治疗方案】: 一些评论人可能会给出进一步治疗的建议, 如果该治疗建议时基于某诊断,
请以"治疗建议 (基于诊断: xxx) "的形式总结
-【检查建议】: 一些评论人可能会给出进一步检查的建议, 如果该检查建议时基于某诊断,
请以" 检查建议 (基于诊断: xxx) "的形式总结

以上【】中的条目不一定在每个帖子中全部存在, 在生成摘要时若某条目不存在,
请以 “【条目】: 无” 的方式写出, 不要省略这些条目。

为帮助你更好生成摘要，我将提供给你你在之前任务重对于该论坛帖抽取出来的相关信息。
请你结合你前期的抽取和帖子本身的内容，生成摘要。
在【评论区诊断】你需要整合之前你对Diagnosis的抽取并计数。

#OBJECTIVE#
现在我将给出论坛帖子，请仔细阅读并总结。
{content}

这是你前一轮所抽取的信息:
{previous_ie}

#STYLE#
你应该尽可能直接从帖子中抽取相关信息生成摘要，最好不要转写。

#TONE#
客观抽取信息，不要给出评论。

#AUDIENCE#
这个摘要可以帮助人们更好去理解论坛帖子，使信息更整洁且有结构。

#RESPONSE#
你需要按照以下格式生成摘要并且利用抽取信息:
【病人信息】: Age, Sex
【病史】: Clinical_History
【前期诊断】: Previous_Diagnosis
【前期治疗】: Previous_Treatment
【发帖人主要诉求】: Problem_Current

【持续时间】: Duration

【评论区诊断】: Diagnosis
xx人xxx
【治疗建议】: Treatment
xxx
【检查建议】: Test
xxx

注意，只需要输出以上摘要，不需要给出任何其他的解释或者评论或任何你思考的过程。

A.6 Few-shot extraction-then-summarization

#CONTEXT#
我希望你帮我总结以下医疗论坛讨论的帖子，抽取相关信息，并生成结构性的摘要。
我将给你提供<post_title>, <post_content>, <comments>,
提供内容中也包含发帖人和评论人的id，有时发帖人也会在评论区回复评论或补充信息。
我希望你从以下几个方面在生成摘要。
-【病人信息】: 包括病人的年龄、性别
-【病史】: 病人曾经的疾病（不包含当前在帖子里询问的特定疾病）
-【前期诊断】: 对当前询问疾病的曾经的诊断
-【前期治疗】: 对当前询问疾病的曾经的治疗，若有前期治疗效果也需要包含
-【发帖人主要诉求】: 本帖主要想询问或讨论的内容（包含对当前疾病情况的描述）
-【持续时间】当前询问的疾病的持续时间
-【评论区诊断】评论区对于发帖人询问的诊断，某些发帖人可能会给出相同的诊断，
并且他们的诊断有确定性的区分（是、可能是、不是、基于...条件是），
所以给出评论区诊断时你应该整理信息，形成"N人认为(确定性xxx)(诊断内容xxx)"这样结构的摘要
-【治疗方案】: 一些评论人可能会给出进一步治疗的建议，如果该治疗建议时基于某诊断，
请以"治疗建议（基于诊断: xxx）"的形式总结

-【检查建议】：一些评论人可能会给出进一步检查的建议，如果该检查建议时基于某诊断，请以”检查建议（基于诊断：xxx）"的形式总结

以上【】中的条目不一定在每个帖子中全部存在，在生成摘要时若某条目不存在，
请以“【条目】：无”的方式写出，不要省略这些条目。

为帮助你更好生成摘要，我将提供给你你在之前任务重对于该论坛帖抽取出来的相关信息。
请你结合你前期的抽取和帖子本身的内容，生成摘要。
在【评论区诊断】你需要整合之前你对Diagnosis的抽取并计数。

你可以参照以下我提供给你的例子学习如何利用抽取的信息生成摘要:
{ie_sum_example}

#OBJECTIVE#
现在我将给出论坛帖子，请仔细阅读并总结。
{content}

这是你前一轮所抽取的信息:
{previous_ie}

#STYLE#
你应该尽可能直接从帖子中抽取相关信息生成摘要，最好不要转写。

#TONE#
客观抽取信息，不要给出评论。

#AUDIENCE#
这个摘要可以帮助人们更好去理解论坛帖子，使信息更整洁且有结构。

#RESPONSE#

你需要按照以下格式生成摘要并且利用抽取信息:

【病人信息】: Age, Sex

【病史】: Clinical_History

【前期诊断】: Previous_Diagnosis

【前期治疗】: Previous_Treatment

【发帖人主要诉求】: Problem_Current

【持续时间】: Duration

【评论区诊断】: Diagnosis

xx人xxx

【治疗建议】: Treatment

xxx

【检查建议】: Test

xxx

注意，只需要输出以上摘要，不需要给出任何其他的解释或者评论或任何你思考的过程。