SummQA at MEDIQA-Chat 2023:
In-Context Learning with GPT-4 for Medical Summarization

Yash Mathur[*] **Sanketh Rangreji**[*] **Raghav Kapoor**[*]
Medha Palavalli[*] **Amanda Bertsch** **Matthew R. Gormley**
Carnegie Mellon University
[ymathur, srangrej, raghavka, mpalaval, abertsch, mgormley] @andrew.cmu.edu

Abstract

Medical dialogue summarization is challenging due to the unstructured nature of medical conversations, the use of medical terminology in gold summaries, and the need to identify key information across multiple symptom sets. We present a novel system for the Dialogue2Note Medical Summarization tasks in the MEDIQA 2023 Shared Task. Our approach for section-wise summarization (Task A) is a two-stage process of selecting semantically similar dialogues and using the top-k similar dialogues as in-context examples for GPT-4. For full-note summarization (Task B), we use a similar solution with k=1. We achieved 3rd place in Task A (2nd among all teams), 4th place in Task B Division Wise Summarization (2nd among all teams), 15th place in Task A Section Header Classification (9th among all teams), and 8th place among all teams in Task B. Our results highlight the effectiveness of few-shot prompting for this task, though we also identify several weaknesses of prompting-based approaches. We compare GPT-4 performance with several finetuned baselines. We find that GPT-4 summaries are more abstractive and shorter. We make our code publicly available [1].

1 Introduction

Medical dialogue summarization is a long-standing challenge in NLP (López Espejel, 2019; Joshi et al., 2020a; Chintagunta et al., 2021; Navarro et al., 2022). Medical scribes write notes on doctor-patient conversations in a predefined template called SOAP notes (Ullman et al., 2021; Podder et al., 2020), which contains sections for information from the patient, test results and observations, diagnosis, and the conclusion or treatment.

Medical summarization is challenging for several reasons. It requires dialogue understanding, where data is often limited (Dai et al., 2020; Lin et al., 2020); this is compounded by the sensitive nature of medical information, which restricts the release of training data for this task (Johnson et al., 2023). Doctors and patients may discuss several conditions in the same conversation, requiring the scribe to differentiate (Gidwani et al., 2017; Mishra et al., 2018). Scribes often use medical terminology in the notes that are not present in the doctor-patient conversation (Corby et al., 2020). Additionally, medical summarization is a high-stakes domain (Naik et al., 2022), motivating several efforts to build explainabile systems for this task (Jain et al., 2022; Reddy, 2022). In parallel, research on large language models (LLMs) has demonstrated compelling few-shot capabilities across domains (Brown et al., 2020; Perez et al., 2021).

In this paper, we explore several potential applications of a recent LLM, GPT-4 (OpenAI, 2023), on medical summarization. We use GPT-4 and finetuned BioBERT (Lee et al., 2020) as an ensemble for classifying the section headers of medical summaries, a 20-category classification problem. Then, given a candidate section header, we apply Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) to select examples for a fewshot demonstration and use these examples to prompt GPT-4 for section-wise summarization. This approach outperforms finetuning BART (Lewis et al., 2019) and T5 (Raffel et al., 2020) over the limited available data. For full-note summarization, we take a similar approach, but select only a single example for the demonstration due to the increased length of the inputs. This also outperforms our supervised baselines. We outline several additional potential prompting approaches and compare their relative efficacy.

Applying LLMs for medical summarization is a compelling solution to the data scarcity problems in this domain, and we find promising performance, with our team placing second in the MEDIQA 2023 Shared Task for Subtask A and Division Summary

---

[*]Equal contribution
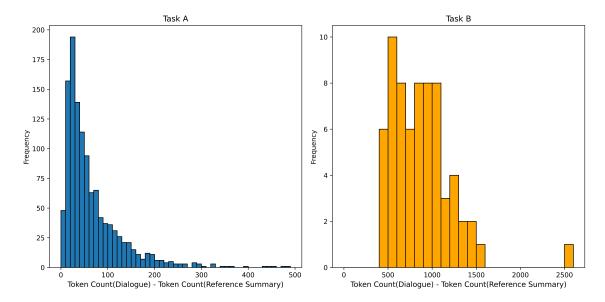[1]https://github.com/Raghav1606/SummQA

Figure 1: Distribution of difference in length between dialogue and reference summary. A larger difference in length indicates a higher degree of compression.
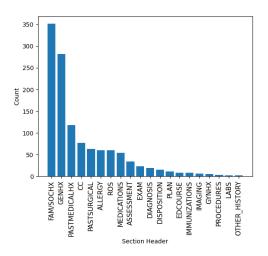

Figure 2: Section header distribution for Task A.

for Subtask B. However, we also identify key areas for improvement. We analyze the differences in outputs between the settings in output length and extractive ability. We find that the summaries generated by LLMs tend to be shorter and less extractive than human-generated summaries as well as SOTA fine-tuned biomedical summarization models. We also note the impracticality of this approach for real data, due to privacy concerns.

2 Background

Dialogue2Note Summarization was one of two tracks in the MEDIQA-Chat 2023 shared task (Ben Abacha et al., 2023). The track was further comprised of two tasks.

Task A involves generating a section-specific clinical summary from a conversation between a patient and a doctor. Additionally, Task A includes a classification task: assigning each dialogue an appropriate section header. There are 1200 conversations in the training split of the dataset (Ben Abacha et al., 2023) for Task A. The distribution over the section headers in Task A is a long-tailed distribution, displayed in Figure 2.

Task B involves generating a full note summary given a conversation; these summaries were evaluated on the section-level and the full-note level. There were 67 conversations in the training split of the dataset (Yim et al., 2023); these dialogues and reference summaries are significantly longer than those for Task A, as these dialogues encompass an entire conversation between a patient and a doctor. The distribution of the difference in dialogue and summary length for both tasks is shown in Figure 1.

3 Related Work

Summarization In recent years, fine-tuning pre-trained models on domain-specific datasets has been the leading practice in text summarization research. While these models produce high-quality summaries and earn high scores against standard benchmarks, they require large datasets in order to adapt to specific domains or summarization styles (Lewis et al., 2020). Transformer-based models (Michalopoulos et al., 2022) and pointer generator
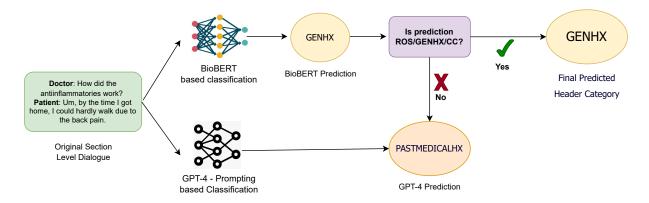
Figure 3: System Architecture for Section Header Classification (Task A)

network models (Joshi et al., 2020b) have been fine-tuned with medical domain knowledge to produce summaries that achieve state-of-the-art results.

Maximal Marginal Relevance was created to reduce redundancies in multi-document summaries (Goldstein and Carbonell, 1998). Abdullah et al. (2023) used MMR to generate query-focused summaries from pre-trained models without performing fine-tuning. (Ye et al., 2022) use MMR to select examples for in-context prompting.

The success of prompt-based models such as GPT-3 (Brown et al., 2020) has allowed for learning from natural language task instructions and/or a few demonstrative examples in the context without updating model parameters. In news summarization, Goyal et al. (2022) find that GPT-3 summaries were preferred by humans over summaries from fine-tuned models trained on large summarization datasets; they posit that zero-shot summaries avoid pitfalls from low-quality training data that are common in summaries from fine-tuned models. In the biomedical domain, pre-trained language models and few-shot learning has been used to collect and generate labeled data for medical dialogue summarization (Chintagunta et al., 2021). Recent work has used GPT-4 to pass the USMLE without any specialized prompt crafting (Nori et al., 2023) and perform zero-shot medical evidence summarization across six clinical domains (Tang et al., 2023).

Few-shot learning Few-shot learning can be unstable as the prompt format, training examples, and even the order of the training examples can cause accuracy to vary from near chance to near state-of-the-art (Lu et al., 2022). Recent work on prompting has tried to mitigate these problems through techniques such as calibration (Zhao et al., 2021), prompt combination (Zhou et al., 2022), or auto-matic prompt generation (Gao et al., 2021).

To mitigate any instability caused by a model's bias, Zhao et al. (2021) estimated the bias towards each answer by asking for its prediction when given the training prompt and a content-free test input such as "N/A" and then fit calibration parameters that cause the prediction for this input to be uniform across answers. To date, studies in prompt combination are rooted in paraphrasing-based methods that take a seed prompt and paraphrase it into several semantically similar expressions. Typically simple ensemble methods (Zhou et al., 2022) such as Maximal Marginal Relevance (Mao et al., 2020) are used to combine the answers to the different prompts as to provide each prompt to contribute to the final answer.

A number of techniques have also been proposed for selecting fewshot examples (Rubin et al., 2022). Fewshot techniques often rely on selecting optimal examples from a large dataset; some work has shown that this leads to an overstatement of few-shot performance, as a large number of labeled examples are necessary to select good examples for the fewshot prompt (Perez et al., 2021). We note that we use the full datasets (1,200 examples for Task A, 40 for Task B) for our prompt selection techniques.

4 Methodology and Baselines

Our summary generation pipeline remains the same across the two tasks: we use GPT-4 to generate a summary given k in-context examples.

4.1 Task A - Section Level Summary

Task A is composed of two subtasks, namely the section header classification and the section-level summarization. We discuss our approach for each
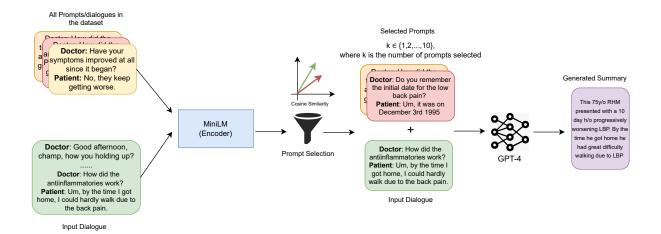
Figure 4: System Architecture for Summarization Task

of the subtasks below.

4.1.1 Section Header Classification

For the section header classification task, we create an ensemble of two models: BioBERT (Lee et al., 2020) and GPT-4. We fine-tune BioBERT with the training data provided for task A. We leverage GPT-4 to perform zero-shot classification on a sample with a given prompt (shown in Table 7). During our analysis of each model's performance, we observe that each model is more accurate than the other on a distinct subset of classes. To leverage the varying nature of predictions from the models we build an ensemble classifier. The overall accuracies are shown in Table 2. We observe empirically that our prompting-based approaches do not perform well on three of the section headers: ROS (Review of Systems), GENHX (History of the present illness), and CC (Chief Complaint), To create an ensemble classifier, we select BioBERT's predictions when it classifies a dialogue as one of these three section headers, and we select the prediction of our GPT-4 based pipeline otherwise. We present the architecture of our final model in Figure 3.

4.1.2 Section Summary

To generate summaries for a given section, we follow a multi-step process as shown in Figure 4. We encode each dialogue in the training data with MiniLM (Wang et al., 2020). For each dialogue to be summarized, we calculate cosine similarities with encoded dialogues from the training data. We retrieve $k=7$ similar examples from the training data based on the highest similarity. This similarity search, using cosine-similarity, serves as a prompt selection method, and the resulting few-

shot prompts, with $k=7$ are then fed to GPT-4 along with their section headers to obtain a summary for the given section. We provide the prompt templates used in Table 7. We select $k=7$ as this fits well in the context length of our prompting-based pipeline; we perform an analysis with varying k in section 5.5.

4.2 In-context Example Selection for Summarization Tasks

This approach involves the dynamic selection of in-context examples for each example during validation or testing. This process entails utilizing matching or similarity criteria to compare the input dialogue of a specific example to a candidate pool comprising the complete training set. Through this process, we are able to select the most suitable examples for each individual case, thereby enhancing the efficacy of our prompts.

Semantic Similarity   Here we select the k most similar examples (dialogue and summary pairs) based on semantic similarity between the provided input dialogue and the input dialogues in the training set. We store the selected examples and generate prompts which we then evaluate on the validation/test set.

Maximal Marginal Relevance   We select k few-shot prompts using Maximal Marginal Relevance (MMR). Similarly to Ye et al. (2022), we use MMR to select an example and use it as a one-shot example for prompting. Our choice of MMR was motivated by the idea that the diversity in the selected in-context examples of the prompt would help with some generalization;.

4.3 Task B - Full Note Summary

For the summarization of entire dialogues, our goal is to generate a full note containing all the appropriate sections. We use a similar approach as described in section 4.1.2 but we restrict it to k=1 similar examples from the training set and include section-level headers in the prompts to help the model understand the sections in the sumary. We selected one in-context example due to long example length relative to the context window of the model. The one-shot prompt is then again fed to the GPT-4 model to obtain a full-length summary. The results f

4.4 Baseline Approaches

We also consider a variety of baseline approaches including, supervised fine-tuning of T5, zero-shot/few-shot GPT-3, perspective-shifting the dialogue followed by summarization, two-stage prompting, our similarity-based in-context learning applied to GPT-3, and mixing of extractive/abstractive methods.

4.4.1 T5

We fine-tuned the T5-small model for the end-end full-length summarization task (Task B). We fine-tuned for 20 epochs with a learning rate of 0.001. Our objective was to obtain a basic model that can serve as a benchmark to assess the complexity and difficulties associated with this specific task. We find that this finetuned model significantly under-performs our other methods, with a ROUGE-1 of 20.187; this may be due to the small dataset for finetuning or a non-optimized set of hyperparameters, as we do not do extensive hyperparameter search.

4.4.2 GPT-3

We investigated several prompting strategies and approaches using *text-davinci-003*.

Zero-shot prompting   For Task B we used the prompt template mentioned in the Appendix A, where we specified the dialogue to be summarized with an instruction prompt mentioning the 4 main sections usually reported in the SOAP notes - "HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "RESULTS" and "ASSESSMENT AND PLAN". The zero-shot prompt gave us a reasonably high ROUGE-1 score of 45.911.

4.4.3 Few-shot prompting for section-wise summary

For Task A, we employed `text-davinci-003` few-shot prompting strategy. Initially, we grouped and categorized the existing 20 section headers for the dataset into 4 main sections, namely "HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "ASSESSMENT AND PLAN", and "RESULTS". The categorization scheme is detailed in Table 6. It is worth noting that "Medications" can be categorized under either "HISTORY OF PRESENT ILLNESS" or "ASSESSMENT AND PLAN" and therefore appears in both categories. We created four few-shot prompt templates, each comprising k=5 in-context examples, for each section. For each example in the validation set, we selected the appropriate prompt based on the classified section header.

Perspective Shift   In this method which we evaluated for Task B, we adopt a two-stage prompting approach where we first use gpt3.5-turbo to obtain a third-person narrative from the input dialogue, following Bertsch et al. (2022), and use the third-person perspective narrative generated as input to a *text-davinci-003* model to generate a summary using the same instruction prompt specifying each section header that needs to be generated.

Two Stage Prompting   In this approach we defined two chained prompts applied one after the other in a stage-wise manner. The first stage prompt was "List the important points from the above conversation for a medical report". This generated a list of salient points summarizing the dialogue. The second stage prompt we used was "Create a paragraph from the above facts only". The output from this prompt served as the final summary, which we then evaluated. We opted for these specific phrasings in the second prompt to mitigate the issue of model hallucination, which we observed was prevalent when tasked with generating a medical summary directly.

5 Results and Analysis

5.1 Experimental Setup

We used an 80/20 train/validation split on the training set and used the entire validation split as our test set. The main hyperparameter that we varied across our experiments for prompt selection was k, the number of in-context examples we selected for the prompt. We report the ablation study on

Task	Models	R1	R2	RL	BR	BP	BF1	BL
	Few-shot Text-davinci-003	17.3	6.8	13.5	0.562	0.539	0.546	0.398
	T5-Small	30.0	11.3	23.6	0.631	0.675	0.646	0.445
A	Two Stage Prompting	28.4	11.6	21.6	0.694	0.633	0.656	0.547
	Prompt Selection text-davinci-003 (Semantic)	38.6	18.4	31.2	0.716	0.725	0.715	0.56
	Prompt Selection text-davinci-003 (MMR)	39.2	**18.8**	31.9	0.717	0.725	0.716	0.559
	Prompt Selection GPT4 (Semantic)	**42.8**	17.2	**32.3**	**0.719**	**0.729**	**0.720**	**0.564**
	T5-Small	22.7	9.1	12.9	0.568	0.471	0.514	0.319
B	Perspective Shift	38.7	15.6	23.8	0.618	0.679	0.647	0.392
	Zero Shot Text-davinci-003	45.4	23.3	30.6	0.644	0.712	0.676	0.410
	Prompt Selection GPT4 (k = 1)	**50.7**	**24.9**	**33.6**	**0.666**	**0.703**	**0.684**	**0.406**

Table 1: Validation Results for Task A and Task B Summarization. Metrics include ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BERTScore Precision (BP), Recall (BR), and F1 (BF1), and BLEURT (BL).

Model	Accuracy
GPT-3.5-turbo	68.943
GPT-4	69.474
BioBERT	71.278
Ensemble (GPT-4 + BioBERT)	**75.312**

Table 2: Validation Results for Header Classification

Task	Summary	EFC	EFD	CR
A	Reference	0.689	1.648	3.387
	Generated	0.561	1.036	5.701
B	Reference	0.671	2.044	2.856
	Generated	0.781	3.086	5.281

Table 3: Summary extractiveness comparison - Extractive Fragment Coverage(EFC), Extractive Fragment Density(EFD), Compression Ratio (CR)

varying k over the validation split in Table 5. For generations we used a single decoding (n = 1), temperature = 1.0, top_p = 1.0 and max_{tokens} = 800. The metrics for BERTScore and BLEURT in Table 1 have been calculated using RoBERTa Large (Liu et al., 2019) and BLEURT-Tiny[2] respectively.

5.2 Experimental Results

Our experiment involving prompt selection via semantic similarity with GPT-4 yielded the most favorable outcomes on the validation split, and prompt selection was the best approach for both Task A and Task B. We propose that the remarkable performance of prompt selection is attributed to the in-context examples that were selected using semantic similarity with the input dialogue. This approach facilitates the generation of an example-specific prompt that incorporates similar in-context examples, leading to an improvement in the model's ability to produce summaries that are more relevant and precise. The use of semantic similarity allows for the identification of examples that share similar semantic structures with the input dialogue, thereby increasing the likelihood of generating coherent and accurate summaries.

5.3 Length of Generated Summary vs. Reference Summary

As shown in Figure 5 we see that most generated summaries were shorter than reference summaries across tasks. This difference was more pronounced in Task B and therefore the summaries produced by our approach fall short in length thereby affecting the ROUGE-1 score as the number of matching n-grams is less. However, we observe that the BERT score still remains consistent even while producing shorter summaries.

Another interesting observation is that individual section summaries, when combined together to produce a full-length summary are closer to the original length rather than prompting GPT-4 to generate a complete summary together. Hence, ensembling multiple section-level summaries to produce a longer summary is an approach we can explore further. We also tried multiple prompt templates (refer Table 7), encouraging the model to produce longer summaries. However, the fact that we require a *summary* induces the model to be concise.

5.4 Extractiveness of Summaries

We measure the extractiveness of the generated summaries using three measures namely. extractive fragment coverage (EFC) (Grusky et al., 2020),
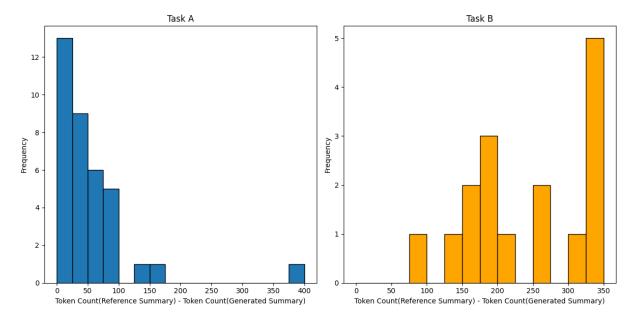
Figure 5: Difference in length of reference summaries and generated summaries

extractive fragment density (EFD) (Grusky et al., 2020) and compression ratio (CR) (Grusky et al., 2020). These metrics rely upon the concept of extractive fragments which are defined as shared sequences in the dialogue and the summary. The extractive fragment coverage quantifies the percentage of words in the summary that are a part of the extractive fragments in the original dialogue. The extractive fragment density measures the average length of the extractive fragment to which a word in the summary belongs to. Compression is measured as the fraction of words in the article and summary.

A comparison of the extractiveness and compression ratio of the reference and generated summaries are shown in Table 3. For Task A, the relatively poor extractive performance of our proposed methods could be due to the smaller size of generated summaries which prevents the usage of more terms from the dialogue. In Task B, we observe the extractive capability of our model improving. This could be attributed to the larger dialogues in Task B allowing for a larger candidate space of tokens to be used in the generations. The generated summaries in Task B are still smaller than the reference summaries as shown by the compression ratio.

5.5 Impact of the number of in-context examples

We further evaluate the impact of the number of in-context examples (k) on various metrics. We report the metrics in Table 5. We observe a general improvement across all metrics as we increase k.

This implies that the generated summaries improve as the model is given more access to relevant data through in-context examples. The relevancy in our method is brought in through the selection of in-context examples via semantic similarity/maximum marginal relevancy. This experiment was only performed for Task A since the token limits of the models did not permit the ablation of k for Task B.

6 Future Work and Conclusion

This paper attempts to automatically generate summaries or structured SOAP notes from a conversation between a doctor and a patient. We tackle this problem by generating section-wise summaries, classifying these summaries into appropriate section headers and generating full-length summaries from longer conversations.

We conclude from the results that prompting-based techniques by themselves can not perform optimally themselves but do give some outstanding results when combined with existing techniques, like prompt selection using MiniLM. Next, we also dive deep into where prompting-based methods underperform the standard models like BART and T5.

Overall, our model concluded 3rd amongst all runs submitted and 2nd as a team for task A, which aimed at producing section-level summaries. Our system also stood 4th amongst all runs submitted and 2nd as a team in division-wise summaries for task B. In the future, we plan to use an ensemble of extractive and abstractive methods of generating

Task	Runs	R1	R2	RL	BR	BP	BF1	BL
	Run 1	42.8	17.2	32.3	0.719	0.729	0.720	0.564
A	Run 2	42.8	17.1	32.4	0.718	0.727	0.717	0.558
	Run 3	42.6	17.1	32.1	0.716	0.724	0.714	0.551
	Run 1	50.7	24.9	33.6	0.666	0.703	0.684	0.406
B	Run 2	51.3	24.9	33.7	0.668	0.704	0.686	0.411
	Run 3	50.9	24.8	33.6	0.666	0.704	0.684	0.407

Table 4: Stability of Validation Results. Metrics include ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BERTScore Precision (BP), Recall (BR), and F1 (BF1), and BLUERT (BL).

k	R1	BP	BR	BF1	BL
3	40.3	0.710	0.693	0.708	0.486
5	41.9	0.718	0.706	0.714	0.521
7	42.8	0.729	0.719	0.720	0.564

Table 5: Impact of number of in-context examples (k) for Task A (GPT-4)

summaries as well as using diversity algorithms that will aid in producing SOAP notes that are more robust and apt as per human evaluations.

7 Limitations

Considering the critical nature of the domain of the task, it is of paramount importance to ensure stability in the results expected from the model. Despite setting the temperature (T) as 0 for all decodings in our experiments, we observe the variance in the generated summaries across runs. Table 4 contains the results for three runs for Task A and Task B. The in-context examples for each sample and other parameters have been kept constant across these runs to identify the degree of stochasticity. Further, in-context learning has shown to be susceptible to changes in order of in-context examples (Lu et al., 2021), as well as the template of the examples (Shin et al., 2020). A more reliable process to generate the summaries along with identification of the optimal examples (template, order) is thus required. Additionally, due to the context limit of the GPT-4 model, evaluating the impact of natural language instructions in addition to the examples could not be performed.

8 Ethics Statement

There exist several risks and ethical considerations that necessitate comprehensive addressal prior to the deployment and utilization of our proposed methods utilizing Language Models (LLMs). A significant apprehension associated with employ-ing LLMs for summarization, as evidenced during experimentation, is their susceptibility to hallucination. This means that there would need to be stringent and effective fact-verification post-processing on the generated summaries, thereby ensuring their factual accuracy and alignment with the doctor-patient discourse.

The preservation of patient confidentiality and privacy assumes paramount importance within the context of healthcare data, given its highly sensitive and personal nature. Consequently, it becomes imperative to undertake effective data anonymization techniques to safeguard patient identities. Additionally, obtaining explicit consent from patients regarding the utilization of their data assumes critical significance. In tandem, strict adherence to the standards set forth by the Health Insurance Portability and Accountability Act (HIPAA) is essential to ensure compliance and guarantee the privacy and security of patient information.

Furthermore, another vital aspect that demands careful consideration is the explainability and interpretability when utilizing Language Models (LLMs) for medical summarization. It becomes essential to address the challenge of comprehending the decision-making processes underlying their outputs. Particularly within the medical domain, where critical decisions are made based on these outputs, explainability is of great importance.

References

Deen Abdullah, Shamanth Nayak, Gandharv Suri, and Yllias Chali. 2023. Generating query focused summaries without fine-tuning the transformer-based pre-trained models.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Amanda Bertsch, Graham Neubig, and Matthew R. Gormley. 2022. He said, she said: Style transfer for shifting the perspective of dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4823–4840, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, and Melanie Subbiah. 2020. Language models are few-shot learners.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Sky Corby, Jeffrey A. Gold, Vishnu Mohan, Nicholas Solberg, James Becton, Robby Bergstrom, Benjamin Orwoll, Christopher Hoekstra, and Joan S. Ash. 2020. A sociotechnical multiple perspectives approach to the use of medical scribes: A deeper dive into the scribe-provider interaction.

Yinpei Dai, Hangyu Li, Chengguang Tang, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 609–618, Online. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen†. 2021. Making pre-trained language models better few-shot learners.

Risha Gidwani, Cathina Nguyen, Alexis Kofoed, Catherine Carragee, Tracy Rydel, Ian Nelligan, Amelia Sattler, Megan Mahoney, and Steven Lin. 2017. Impact of scribes on physician satisfaction, patient satisfaction, and charting efficiency: A randomized controlled trial. *Annals of family medicine*, 15:427–433.

Jade Goldstein and Jaime Carbonell. 1998. Summarization: (1) using MMR for diversity- based reranking and (2) evaluating summaries. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 181–195, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies.

Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2).

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020a. Dr. summarize: Global summarization of medical dialogue by exploiting local structures.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020b. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2020. Graph-evolving meta-learning for low-resource medical dialogue generation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jessica López Espejel. 2019. Automatic summarization of medical conversations, a review. In *Actes de la Conférence sur le Traitement Automatique*

des Langues Naturelles (TALN) PFIA 2019. Volume III : RECITAL, pages 487–498, Toulouse, France. ATALA.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1737–1751, Online. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pranita Mishra, Jacqueline C. Kiang, and Richard W. Grant. 2018. Association of medical scribes in primary care with physician workflow and patient experience. *JAMA internal medicine*, 178:1467–1472.

Nithesh Naik, B. M. Zeeshan Hameed, Dasharathraj K. Shetty, Dishant Swain, Milap Shah, Rahul Paul, Kaivalya Aggarwal, Sufyan Ibrahim, Vathsala Patil, Komal Smriti, Suyog Shetty, Bhavan Prasad Rai, Piotr Chlosta, and Bhaskar K. Somani. 2022. Legal and ethical consideration in artificial intelligence in healthcare: Who takes responsibility? *Frontiers in surgery*, 9.

David Fraile Navarro, Mark Dras, and Shlomo Berkovsky. 2022. Few-shot fine-tuning SOTA summarization models for medical dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 254–266, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

OpenAI. 2023. Gpt-4 technical report.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2020. Soap notes.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sandeep Reddy. 2022. Explainability and artificial intelligence in medicine. the lancet. digital health.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *medRxiv*.

Kristen Ullman, Lauren McKenzie, Bradley Bart, Glennon Park, Roderick MacDonald, Eric Linskens, and Timothy J. Wilt. 2021. The effect of medical scribes in emergency departments: A systematic review. *The Journal of Emergency Medicine*, 61:19–28.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Prompt consistency for zero-shot task generalization.

A Appendix

This appendix presents two tables - Table 6 contains the categories and subcategories in which the dialogue is divided to create a SOAP note. Table 7 presents the prompts used by approaches for tasks A and B.

Table 6: Categorization Scheme

HISTORY OF PRESENT ILLNESS
Fam/Sochx [Family History/Social History]
Genhx [History of Present Illness]
Pastmedicalhx [Past Medical History]
CC [Chief Complain]
Pastsurgical [Past Surgical History]
Allergy
Gynhx [Gynecologic History]
Other_history
Immunizations
Medications
PHYSICAL EXAM
ROS [Review of Systems]
Exam
RESULTS
Imaging
Procedures
Labs
ASSESSMENT AND PLAN
Assessment
Diagnosis
Plan
Edcourse [Emergency Department Course]
Disposition
Medications

Table 7: Prompt Templates

Prompting Approach	Model	Prompt(Example)
Zero-Shot	text-davinci-003	"Summarize the following into a medical report having the following sections: 'HISTORY OF PRESENT ILLNESS', 'PHYSICAL EXAM', 'RESULTS', 'ASSESSMENT AND PLAN'.
Few-shot prompting for section-wise summary	text-davinci-003	Prompt for PHYSICAL EXAM section(k=5) """ Dialogue: Doctor: Breath in breath out, let me tap it and see. Well, your lungs sound clear. Patient: Okay. Summary: CHEST: Lungs bilaterally clear to auscultation and percussion. ….. ….. ….. Dialogue : Doctor: Do you have any chest pain? Patient: No, I don't. Doctor: Any breathlessness? Patient: Yes, I do get breathless only when I have to do some form of exertion like walking a long time or running. Doctor: Okay. How about any bowel issues? Patient: No, I don't have any stomach problems except I have to go frequently to use the bathroom. Doctor: Okay frequency. How about any prolonged bleeding issues or anything like that sort? Patient: No nothing like that. Summary : He denies any chest pain. He admits to exertional shortness of breath. He denies any GI problems as noted. Has frequent urination as noted. He denies any bleeding disorders or bleeding history. Dialogue : {dialogue} Summary : """
Perspective Shift	text-davinci-003 gpt3.5-turbo	2 staged prompting (perspective shift with turbo and summarization with davinci) PERSPECTIVE SHIFT = """ Convert the following into third person. {dialogue} \\\\\\ """ PROMPT = """ Summarize the following into a medical report having the following sections: "HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "RESULTS", "ASSESSMENT AND PLAN" where each section is at least 60 words. {third-person-perspective}\\\\\\ """
Two-Stage Prompting	text-davinci-003	PROMPT #1 """ {dialogue} \\ List the important points from the above conversation for a medical report """ PROMPT #2 """ {prompt1-generated-output} \\ Create a paragraph from the above facts only """

Prompting Approach	Model	Prompt(Example)
Prompt Selection - MMR(k=3)	text-davinci-003	PROMPT SELECTION with k=3 Dialogue: Doctor: Your last visit was on April seventh two thousand five, correct? Patient: Ah no, it was on April eighth two thousand five, doctor. Doctor: That's right. So, has anything changed since then? Patient: No, everything is the same really. Summary: Essentially unchanged from my visit of 04/08/2005. ….. ….. ….. Dialogue: Doctor: Do you have any past or present medical conditions? Patient: No. Summary: None. ….. ….. ….. Dialogue: {dialogue} Summary:
Prompt Selection - Semantic Similarity(k=7) - Task A	**GPT-4**	PROMPT SELECTION with k=7 Dialogue: Doctor: Do you know about any medical issues running in your family? Patient: Yeah, almost everyone had diabetes. Summary: Multiple family members have diabetes mellitus. ….. ….. ….. Dialogue: Doctor: Any specific family medical history that I should be aware of? Patient: No. Doctor: Anyone in your family, even grandparents, if you know them, did they have diabetes or high blood pressure? Patient: No. Doctor: Anyone else sick at home? Patient: No. Summary: Noncontributory. No one else at home is sick. ….. ….. ….. Dialogue: {dialogue} Summary:
Prompt Selection - Semantic Similarity(k=1) Task B	**GPT-4**	PROMPT SELECTION with k=1 Dialogue: [doctor] and why is she here ? annual exam. okay. all right. hi, Sarah. how are you ? [patient] good . how are you ? [doctor] i'm good . are you ready to get started ? [patient] yes , i am . [doctor] okay . so Sarah is a 27-year-old female here for her annual visit. So, Sarah, how have you been since the last time I saw you ? ……. ……. Summary: CHIEF COMPLAINT Annual visit. HISTORY OF PRESENT ILLNESS The patient is a 27-year-old female who presents for her annual visit. She reports that she has been struggling with her depression off and on for the past year...... ……. Dialogue: {dialogue} Summary: