

Summarizing online patient conversations using generative language models

Rakhi Asokkumar Subjagouri Nair, Matthias Hartung, Philipp Heinisch, Janik Jaskolski, Cornelius Starke-Knäusel, Susana Veríssimo, Philipp Cimiano, David Maria Schmidt

Submitted to: JMIR Medical Informatics
on: June 04, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Summarizing online patient conversations using generative language models

Rakhi Asokkumar Subjagouri Nair¹; Matthias Hartung²; Philipp Heinisch¹; Janik Jaskolski²; Cornelius Starke-Knäusel²; Susana Veríssimo²; Philipp Cimiano^{1,2}; David Maria Schmidt¹ MSci

¹Bielefeld University Bielefeld DE

²Semalytix Bielefeld DE

Corresponding Author:

Philipp Cimiano
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Bielefeld
DE

Abstract

Background: The World Wide Web, and online communities in particular, play an important role in providing support for patients in living with their condition. Methodologies for patient listening on social media have been proposed. Social media is acknowledged by regulatory bodies (e.g., FDA) as an important source of patient experience data to learn about patients' unmet needs, priorities, and desired outcomes. Yet, most automatic methods to extract patient experience data from online sources are quantitative, without providing deeper qualitative insights. Thus, there is a lack of methods that support the generation of insights at large scale by automatically summarizing patient experiences shared in online fora.

Objective: The objective of this study was to evaluate to what extent state-of-the-art large language models can be used to appropriately summarize posts shared by patients in online fora and health communities. The focus was on the experience of important disease burdens as well as with existing treatments. The study was conducted on posts from patients from 5 online fora related to breast cancer. The goal was in particular to compare the performance of different language models and prompting strategies, and investigate the feasibility of large language models as a way to generate qualitative insights about the patient experience.

Methods: We applied three different language models (FlanT5, GPT-3 and GPT-3.5) on the task of summarizing posts from patients in online communities. The generated summaries were evaluated with respect to 124 manually created summaries as reference, using standard metrics for the evaluation of text generation methods. We tested in particular the effect of fine-tuning and 7 different strategies for prompting (Zero-shot, One-shot, Three-shot, Zero-Shot Directional Stimulus Prompting (DSP), One-shot DSP, and Three-shot DSP, Chain of Thought). As evaluation metrics, we use ROUGE and BERTScore to compare the automatically generated summaries to manually created reference summaries.

Results: Among the zero-shot-based large language models investigated, GPT-3.5 performs better than the other models with respect to the ROUGE metrics as well as with respect to BERTScore. While Zero-shot seems to be a good prompting strategy, overall GPT-3.5 DSP Three-shot had the best results with respect to the above mentioned metrics.. A manual investigation of the summarization of the best-performing method showed that the generated summaries are accurate and plausible compared to the manual summaries.

Conclusions: Our results suggest that state-of-the-art pre-trained large language models can be successfully used to summarize comments shared by patients in online communities to shed light on how patients experience their condition. Future work should investigate the problem of hallucinations and develop approaches to mitigate them. Future work could further investigate to what extent the proposed methods could be used to support the development and mapping of patient journeys.

(JMIR Preprints 04/06/2024:62909)

DOI: <https://doi.org/10.2196/preprints.62909>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/62909>

Original Manuscript

Summarizing online patient conversations using generative language models

Rakhi Asokkumar Subjagouri Nair (Bielefeld University, 0009-0001-8727-2599)

Matthias Hartung (Semalytix, 0000-0001-6549-066X)

Philipp Heinisch (Bielefeld University, 0009-0002-8079-5570)

Janik Jaskolski (Semalytix)

David Maria Schmidt (Bielefeld University, 0000-0001-7728-2884)

Cornelius Starke-Knäusel (Semalytix, 0009-0008-7331-9668)

Susana Veríssimo (Semalytix)

Philipp Cimiano (Bielefeld University, Semalytix, 0000-0002-4771-441X)

Abstract

Background: The World Wide Web, and online communities in particular, play an important role in providing support for patients in coping with their condition. Methodologies for patient listening on social media have been proposed. Social media is acknowledged by regulatory bodies (e.g., by the Food and Drug Administration, FDA) as an important source of patient experience data to learn about patients' unmet needs, priorities, and preferences. Yet, most automatic methods to extract patient experience data from online sources are quantitative, without providing deeper qualitative insights. *Although current generative models already excel at generating syntactically correct and (at first glance) convincing texts, they still tend to struggle with conciseness and robustness, which is crucial for summarizing patient experiences in a useful way.* The open research question we address in this paper is whether current state-of-the-art large language models can automatically summarize text shared by patients in online communities to support the extraction of patient insights at large scale to inform decision making processes in drug development, evidence generation and clinical trial design.

Objective: The objective of this study was to evaluate to what extent state-of-the-art large language models can be used to appropriately summarize posts shared by patients in online fora and health communities. In contrast to most related work, this study concentrates less on the extraction of key information and statistical summaries but more on a qualitative summary of patient experiences. The focus was on the experience of important disease burdens as well as with existing treatments. The study was conducted on posts from patients from 5 online fora related to breast cancer. The study goal was to compare the performance of different language models and prompting strategies, investigating the suitability of large language models as a way to generate qualitative insights about the patient experience. An expected outcome of this study was that size and recency of the model positively impact the summarization performance, whereas prompting techniques were expected to have smaller impacts in comparison.

Methods: In our experimental and comparative study, we applied three different language models (Flan-T5, GPT-3 and GPT-3.5) on the task of summarizing posts from patients in online communities. The generated summaries were evaluated with respect to 124 manually created summaries as reference. We tested in particular the effect of fine-tuning and 7 different strategies for prompting (Zero-shot, One-shot, Three-shot, Zero-Shot Directional Stimulus Prompting (DSP), One-shot DSP, and Three-shot DSP, Chain of Thought). As evaluation metrics, we use two standard metrics from the field of text generation, namely ROUGE and BERTScore to compare the automatically generated summaries to manually created reference summaries. These metrics have been selected as they are standard metrics to compare the quality of automatically generated texts with respect to a gold standard.

Results: Among the zero-shot-based large language models investigated, GPT-3.5 performs better than the other models with respect to the ROUGE metrics (ROUGE-1: 34.00 vs. 32.20 vs. 18.91, ROUGE-2: 10.34 vs. 9.82 vs. 3.29, ROUGE-L: 30.04 vs. 29.10 vs. 17.32 for GPT-3.5, GPT-3 and Flan-T5, respectively) as well as with respect to BERTScore (88.21 vs. 87.86 vs. 83.34 for GPT-3.5, GPT-3 and Flan-T5, respectively). While Zero-shot seems to be a good prompting strategy, overall GPT-3.5 DSP Three-shot had the best results with respect to the above mentioned metrics. A manual investigation of the summarization of the best-performing method showed that the generated summaries are accurate and plausible compared to the manual summaries.

Conclusions: Our results suggest that state-of-the-art pre-trained large language models can be successfully used to summarize comments shared by patients in online communities to shed light on how patients experience their condition. Our experiments show promising results for all state-of-the-art models, with GPT-3.5 performing best. Notable limitations of this study are the relatively small sample size of the manually-created dataset as well as the reliability of the applied automatic text generation evaluation metrics. The key limitations of our work are the limited data sample we rely on as well as the fact that the summaries have been created by one annotator only. Further, the results hold only for the examined models and prompting strategies, potentially not generalizing to other models and strategies. Taken together, our results suggest that state-of-the-art pre-trained language models can be a valuable tool to provide qualitative insights about patient experience to better understand unmet needs, patient priorities, how a disease impacts daily functioning and Quality of Life in order to inform processes aimed at improving healthcare delivery and ensure that drug development focuses more on the actual priorities and unmet needs of patients. In particular, the insights would allow to focus drug development and clinical development on outcomes that are of relevant to patients and have the potential to significantly improve their Quality of Life.

Keywords: patient experience, online communities, summarizing, large language models

Introduction

For many patients, the World Wide Web plays an important supporting role in coping with their condition. In fact, social media sites, digital support groups and fora as well as social networks play an important role in the life of patients, as sharing and consuming experiences of peers provides emotional, informational and psychological support [34]. It has been indeed shown that sharing health-related experiences in online communities has an effect of empowerment on patients, increasing

their subjective and psychological well-being, leading to increased self-management and control [34].

Beyond receiving support, social media is increasingly utilized by patients to educate themselves on a disease, and to find hospitals or physicians most capable of treating their condition [12]. When engaging with social media, users are willing to disclose very intimate information [35]. In fact, it has been shown that the anonymity of many social media sites increases the level of self-disclosure [36]. The consumer survey by PWC [PWC Health Research Institute 2012] has indeed shown that more than 30% of respondents would be comfortable having their social media conversations monitored if that data could help identify ways to improve their health. In a study with adults presenting to an academic, urban emergency department, Padrez et al. (2016) found out that, among patients with a social media profile, 71% consented to share their social media data to compare it to their Electronic Medical Records (EMRs). These figures convey that a significant proportion of patients are willing to share their social media data for research purposes if it contributes to improving their condition and those of their peers. A recent study has confirmed these earlier results, showing that participants would be willing to donate some of their digital data to researchers and clinicians in pursuit of health-related insights [37].

Against this background – the world wide web and social media playing an empowering role in patients' lives, and at the same time patients being willing to donate their data if this contributes to improving their and others' health condition – an important question is how relevant insights can be obtained from the data. The Federal Drug Administration (FDA) has identified social media data as one important source of "patient experience data", i.e. data that can foster understanding of the unmet needs and priorities of patients as a basis to design more effective clinical studies, support outcomes research activities, and ultimately develop better drugs that actually improve patients' lives, well-being and functioning [45]. In fact, methodologies for patient listening are quite mature and are widely used in patient-focused drug development to automatically analyze social media content as a basis to understand which symptoms are most burdensome to patients, how they experience existing treatments, which outcomes matter to them, and overall how their quality of life is affected by the disease [28].

However, these methodologies mostly provide quantitative insights, as they rely on the automatic analysis of social media content using techniques from Natural Language Processing (NLP), extracting key information and summarizing it statistically to get an understanding of trends at the level of an entire patient population or patient cohort of interest. A frequent problem is how to get a more qualitative understanding of how the disease affects patients' daily functioning and living. Most NLP methods in fact can extract single entities or concepts from textual data, but there is a lack of methods to analyze textual data at a more qualitative level in terms of the key messages that can be derived from a collection of texts. Summarization techniques, especially from abstractive summarization, can be used for this purpose, but so far research on how well posts from patients in online communities can be faithfully summarized is missing. While it is possible to manually analyze sample posts, there is the risk of missing relevant aspects, and the task of summarizing all the content manually is infeasible. Thus, there is a need for more scalable methods to derive qualitative insights from online patient experience data.

Considering recent developments in the field of large language models (LLMs), in this article we ask the question to what extent existing pre-trained LLMs can be used to aid the analysis and summa-

rization of patient-authored social media content. Large language models are machine-learned models that have been “pre-trained” on large amounts of text [17] and later either fine-tuned for specific tasks [17-18, 24] or trained to determine a suitable response to a certain instruction. The latter has been referred to as “alignment” and is typically done with reinforcement learning [19]. Overall, large language models have been shown to perform very well on tasks such as summarization [20-21], sentiment analysis [22], question answering [23], and machine translation [24]. While excelling at many tasks, LLMs have also been shown to lack robustness and react too sensitively to small perturbations in the input and to be misled by so called “adversarial examples” [50]. Further, they have been shown to extrapolate too creatively in some cases, generating output that is factually wrong. This phenomenon has been sometimes named as “hallucination” [51].

Based on the need to have advanced methods to summarize the patient experience to yield qualitative insights and the promise of LLMs for many tasks, in this paper we explicitly address the following question: *Can state-of-the-art LLMs be used to accurately summarize the experience of patients as shared in online data sources?*

For this purpose, we experimentally compare different pre-trained large-language models: two from the GPT family (GPT3 and GPT3.5) and Flan-T5. While our focus is on using pre-trained language models that are prompted via instructions, we also carry out fine-tuning experiments to examine the impact of fine-tuning. We experimentally investigate several prompting strategies: Zero-shot, One-shot, Three-shot, both alone as well as in combination with Directional Stimulus Prompting (DSP) and Chain of Thought (CoT) prompting.

For this study, we focused on breast cancer and collected self-reported patient experience data from 5 fora, grouping the documents according to mentions of key symptoms and treatments. The data collected comprises user posts from 5 leading web fora in which breast cancer patients discuss their experiences with peers. The 5 sources have been manually identified in a search process with the aim to identify the most relevant breast cancer fora.

Following recommendations on ethical methods to analyze data using natural language processing [47], the data was automatically anonymized by removing user IDs, personal names, addresses, telephone numbers to minimize the risk of identification of a natural person in order to ensure compliance with GDPR. While the collected data still represents personal information, given the fact that we only analyze data that has been made manifestly public by users without further provisions to protect the data, the use for research purposes can be argued to be acceptable, given that there is a public interest in using this data to improve healthcare and other means to collect such insights (e.g., surveys) would represent a higher burden for patients [47]. On the other side, the harm caused to patients by analyzing the data can be argued to be minimal, as the content is already publicly available and no new information is released as part of the process.

The texts generated by the large language models were evaluated with respect to a dataset of 127 manually produced reference summaries that have been provided by one of the authors with years of expertise in analyzing patient experience data from social media. The generated texts are evaluated in terms of standard text generation metrics including ROUGE metrics and BERTScore. These metrics essentially compute the overlap between the automatically generated texts and the manually created references and thus provide an estimate of the quality of the summaries. In a case study, we discuss the output of different LLMs to get a sense of how these models are actually able to summarize the conversation.

Our results show that advanced prompting techniques significantly improve the ability of generative language models to summarize online patient comments efficiently.

Methods

The research design of this study corresponds to an experimental and comparative study. It corresponds to an experimental study in that it describes experiments in which the ability of LLMs to adequately summarize patient posts was experimentally evaluated by comparing automatically generated summaries to manually created summaries in a series of experiments. The study is also comparative as it evaluates the performance of different LLMs and prompting strategies in comparison to each other. The statistical validity of results is assessed by computing the statistical significance of results by using a t-test, reporting p values comparing the significance of the results of the best performing approach compared to the others. While in most settings we rely on off-the-shelf pre-trained large language models, we also test the effect of fine-tuning on the task.

Dataset and Task Description

For this experimental and comparative study, data was crawled from different online fora in which breast cancer patients discuss their disease experience with peers. The data was collected from the following public fora: breastcancer.org, healthunlocked.com, inspire.com, her2support.org, and macmillan.org.uk based on a search by the authors for the most prominent fora related to breast cancer available on the Web. Data was automatically obtained via SocialGist as a data provider and automatically analyzed to identify posts of patients who have self-disclosed themselves as being breast cancer patients using the methodology described previously by Spies et al. [38]. The preprocessing consists of anonymization, entity extraction and linking via a knowledge graph, detection of symptoms, treatments, quality of life mentions, detection of severity of symptoms as well as detection of impact of symptoms on the quality of life of patients. Also demographic information involving age and gender was extracted if explicitly mentioned in posts.

This yielded a dataset comprising 145,460 documents from 1,146 patients. The preprocessing revealed the most frequent symptoms (8 symptoms) and most frequent treatments used (119). They are listed in Table 1. The data originates from a time frame between August 2016 and July 2022. As expected from a breast cancer population, 99.5% of the patients were female (also male individuals can develop breast cancer in rare cases).

Table 1: List of frequent symptoms and treatments mentioned in the dataset, numbered from 1 to 127

Symptoms	Treatments
----------	------------

1. Rash	9. Abemaciclib, 10. Abraxane, 11. Anastrozole
2. Breast pain	12. Carboplatin, 13. Clinical nutrition, 14. Capecitabine,
3. Nipple discharge	15. Capivasertib 16. Cryosurgery, 17. Cyclophosphamide,
4. Seborrhoeic dermatitis	18. Cytosan, 19. Chemotherapy, 20. Acupuncture,
5. Inverted nipple	21. Adriamycin, 22. Afinitor, 23. Alpelisib, 24. Aranesp,
6. Erythema	25. Aredia, 26. Arimidex, 27. Aromasin, 28. Aromatherapy,
7. Edema	29. Atezolizumab, 30. Avastin, 31. Bevacizumab, 32. Breast reconstruction,
8. Neoplasm	33. Denosumab, 34. Docetaxel, 35. Doxil,
	36. Doxorubicin, 37. Elacestrant, 38. Enhertu, 39. Epirubicin,
	40. Eribulin, 41. Everolimus, 42. Evista, 43. Exemestane,
	44. Exercise, 45. Fareston, 46. Faslodex, 47. Femara,
	48. Filgrastim, 49. Fluorouracil, 50. Fulvestrant,
	51. Gemzar, 52. Giredestrant, 53. Goserelin,
	54. Halaven, 55. Herceptin, 56. Hormonal therapy(oncology),
	57. Hyaluronidase, 58. Hypnotherapy, 59. Ibandronic acid,
	60. Ibrance, 61. Ipatasertib, 62. Ixempra,
	63. Kadcyla, 64. Kanjinti, 65. Keytruda, 66. Kisqali,
	67. Lapatinib, 68. Larotrectinib, 69. Lasofoxifene, 70. Letrozole,
	71. Leuprorelin, 72. Lumpectomy, 73. Lupron,
	74. Lymphadenectomy, 75. Lynparza
	76. Margetuximab, 77. Massage, 78. Mastectomy, 79. Medical Cannabis,
	80. Meditation, 81. Megace, 82. Megestrol,
	83. Mitoxantrone, 84. Navelbine, 85. Neratinib, 86. Nerlynx,
	87. Neulasta, 88. Neupogen, 89. Nivolumab,
	90. Olaparib, 91. Oophorectomy, 92. Opdivo,
	93. Paclitaxel, 94. Palbociclib, 95. Peer support,
	96. Pembrolizumab, 97. Perjeta, 98. Pertuzumab, 99. Phesgo,
	100. Pigray, 101. Radiation therapy, 102. Raloxifene,
	103. Ribociclib, 104. Rintodestrant, 105. Sacituzumab govitecan,
	106. Talazoparib, 107. Talzenna, 108. Tamoxifen, 109. Targeted therapy,
	110. Taxol, 111. Taxotere, 112. Tecentriq,
	113. Toremifene, 114. Trastuzumab, 115. Trastuzumab deruxtecan,
	116. Trodelvy, 117. Tucatinib, 118. Tukysa,
	119. Veliparib, 120. Verzenio, 121. Vinorelbine
	122. Xeloda, 123. Xgeva, 124. Zarxio, 125. Zoladex,
	126. Zoledronic acid, 127. Zometa

Before analysis, the data was automatically anonymized to remove mentions of person names, email, URLs, telephone numbers and addresses. A sample analysis showed that accuracy of anonymization was at 99.5%. For this purpose, we relied on the anonymization solution provided by Private AI¹.

The summarization corpus underlying this study is constructed as follows: For each topic in the dataset (where “topic” refers to one of the key symptoms of breast cancer or one of the existing breast cancer treatments, as listed in Table 1), all documents containing at least one mention of the respective topic are collapsed into a topic-specific sub-corpus. If multiple topics are mentioned within a document, it becomes part of several sub-corpora accordingly. This resulted in 127 topic-

¹ <https://private-ai.com/>

specific sub-corpora, derived from 8 key symptoms and 119 treatments. The number of documents per topic-specific sub-corpus varies between 1 and 44, with a median of 12.

The summarization task can be seen as an instance of multi-document summarization, as it consists in generating a concise summary for each of the 127 topic-specific sub-corpora in the summarization corpus. The main challenges for multi-document summarization are, besides the fact that the input length can be very long, the fact that the content of the documents to summarize might be overlapping, complementary or even inconsistent in the worst case [48].

As a gold standard for reference, one co-author of this paper, SV, who is trained in the annotation of online patient conversation, provided a manual summary of each topic-specific document collection in the summarization corpus. With respect to symptom-related document collections, the annotator was instructed to focus on summarizing the impact of symptoms on patients' daily functioning and Quality of Life (QoL). For treatments, the goal was to summarize outcomes of the treatment, as well as side effects or any other aspect related to the experience with these treatments. Consequently, the gold standard comprises 127 manual summaries for the symptoms and treatments listed in Table 1. Given the significant effort in creating such summaries (around 3.5 hours per summary, close to more than 400 hours in total), it was not feasible to summarize each set of documents by several annotators. We discuss this further in the discussion section while highlighting limitations.

The manual summary for the Rash symptom, for instance, is as follows:

"Patients perceive rashes as being horrible and uncomfortable, yet common. They are a great impairment for many patients since they are often painful and itchy. Some patients report that they cover multiple parts of the body, while others report it being localized to body parts, such as the face or the breast. They also report on the color and feel of the rash: pink, red, photosensitive, raised, large, hive-like, warm, etc. The appearance of the rashes changes with time, for some patients it became better (e.g. less itching), while for others worse (e.g. redder and hotter). Rashes can have different causes: they can be the side effect of treatments, such as chemotherapy, allergic reactions or symptoms. The cause for the rash is often unclear among patients, but it is often said that usually a rash is just a rash."

In order for the texts to fit the input size of the language models used in the study, the length was truncated to 1024 tokens. From the 127 manually generated reference summaries, 3 were used as examples for few-shot training, while the 124 remaining ones were used for evaluation only. For the fine-tuning experiments, the dataset was divided into three non-overlapping sections: train, validation, and test. The resulting data sizes followed an 80:10:10 ratio, resulting in 101, 12, and 14 examples for the train, validation, and test portions, respectively.

Models used for patient comment summarization

For the task of summarizing texts, we rely on pretrained LLMs. On the one hand, we rely on encoder-decoder architectures (Flan-T5), as well as on decoder-only architectures (GPT-models). As our focus lies on the experimental testing and comparison of different prompting techniques (instructions in the form of natural language), we use LLMs which have the capability to follow textual instructions. Flan-T5 and the family of GPT-3 models have this property as they have been trained on data to generate answers for instruction-like prompts using reinforcement learning, a process

known as alignment [49]. We rely on the following three models in particular: Flan-T5 base, GPT-3, and GPT-3.5. Flan-T5 is an enhanced version of T5 that has been fine-tuned on a mixture of tasks using the Flan collection of datasets [1]. For the study, the Flan-T5 base model was used along with GPT-3 and GPT-3.5. GPT-3 is an advanced iteration of the Generative Pre-trained Transformer model, which operates as an autoregressive language model featuring an impressive 175 billion parameters [3]. GPT-3.5 is a refined version of its predecessor, and boasts an enhanced ability to adapt to diverse prompts and instructions, making it a versatile tool for various applications. Together, these models are a good cross-section of the current LLM landscape and well-suited for our task.

Prompting Techniques for Summarization

In order to explore the capabilities of LLMs to summarize patient comments, we compare various prompting strategies. We start with a prompt without examples or guidance on how to solve the task, called **Zero-Shot prompting**. As further methods, we enhance the prompt by introducing examples of summarized patients comments (**Few-Shot prompting**), giving hints (**Directional Stimulus Prompting**), and, finally, guiding the model by asking for intermediate results invoked by a set of self-defined questions (**Chain-of-Thought Prompting**). By combining different techniques (Few-Shot prompting with Directional Stimulus Prompting, and Zero-Shot prompting with Chain-of-Thought Prompting), we investigate the most effective prompting technique for summarizing patients' comments.

One-shot prompting (1SP) closely resembles the few-shot approach, differing in that only a single demonstration is permitted along with the natural language task description [3]. The corresponding used prompt template can be found in Figure 4 in the appendix.

To facilitate our experiment, we adopted this one-shot prompt template. Within the template, `{text}` signifies the patient comments/experience regarding a symptom or treatment drawn from the example set, and `{summary}` represents the corresponding reference summary. The appendix details other prompt templates we exploit in our study.

Few-shot prompting (FSP) introduces k examples consisting of a text and a summary provided by a human expert, followed by the final text which should be summarized by the model. For the study, k is set to 3 due to token limit constraints [3].

Directional Stimulus Prompting (DSP) is a framework that employs an adjustable language model (LM) to offer guidance to a black-box, frozen large language model (LLM) to enhance its performance. Specifically, a policy LM is trained to generate discrete tokens acting as directional stimuli for each input of the LLM. These stimuli, such as keywords extracted from an article for summarization, are then integrated into the original input and fed into the LLM, steering its generation towards the intended target. DSP utilizes a customizable policy LM to produce the stimulus—in the case of summarization, keywords—to direct the LLM in generating a desired summary with improved metrics like higher ROUGE scores or aligning more closely with human preferences. The policy LM can undergo training through supervised fine-tuning from annotated data (SFT) and reinforcement learning from offline and online rewards (RL), allowing for the exploration of directional stimuli that better

align LLMs with human preferences. This adaptable framework is applicable to various language models and tasks [5]. For this study, instead of a policy LM, keyBERT [8] is used with the default parameters along with KeyphraseVectorizers to extract accurate keywords for each patient comment which are given as hints to the summarizing LLM. KeyBERT is a straightforward and user-friendly approach to extracting keywords, utilizing BERT embeddings to generate keywords and keyphrases that closely match the important content of a document [8]. Combining keyphrase vectorizers with KeyBERT for keyphrase extraction leads to a PatternRank approach, enabling the extraction of grammatically accurate keyphrases that closely align with the document's content. In this process, the vectorizer initially extracts candidate keyphrases from text documents, and KeyBERT then ranks these candidates based on their similarity to the document. The top-n most similar keyphrases can be regarded as document keywords. The use of KeyphraseVectorizers alongside KeyBERT offers the advantage of obtaining grammatically correct keyphrases, as opposed to simple n-grams of predefined lengths [9]. In this study, both keywords extracted by keyBERT with KeyphraseVectorizers, in addition to frequent terms occurring in the patient comment, are combined to form a single set which was used as hints/directional stimuli for each patient comment to generate summaries. For example, the hints retrieved and used for the Rash symptom is given as follows: *severe rash; hives rash; rash hives; simple rash; red rash; cause; have; describe; impair; increase*

Chain-of-Thought (CoT) prompting fosters reasoning capabilities by incorporating intermediate reasoning steps. The fundamental concept of this method is straightforward: By introducing phrases like "Let's think step by step" or a similar instruction, it triggers the model to reason step by step, outputting intermediate results that are combined with other intermediate results into an overall summary. In our study, the Chain-of-Thought prompting for the patient comment summary generation was done based on the Element-aware Summarization approach in the Zero-Shot setting. In the element-aware summarization approach using CoT, as part of element extraction, we first ask the LLMs to extract important elements mentioned in the patient comment by manually-set guiding questions given below:

1. What is the main health or medical related topic discussed in this patient comment?
2. What are the important topics mentioned in this patient comment?
3. What are the symptoms and diseases mentioned in this patient comment?
4. What are the recovery methods or treatments mentioned in this patient comment?

Next, we concatenate the extracted elements, i.e., the answers to the above four questions (A) with the prompt (P) "Let's integrate the above information and summarize this patient comment:" along with the patient comment (C), to prompt the LLM for patient comment summary generation. The input to the LLMs is [A; P; C] and the output is the final summary.

The advanced prompting techniques employed in the study were applied in combination with each instruction-based model. Here, we tailor the usage of these prompting techniques to the specific characteristics and insights derived from the data and task setup. For instance, with Directional Stimulus Prompting, keywords extracted from patient comments were used as hints for generating summaries. Similarly, in Chain-of-Thought prompting, guiding questions were manually set to extract important elements from patient comments for summary generation. The prompt templates for summary generation, initially published by Google Research for Flan-T5 [13], were adopted as the baseline prompt template for implementing Zero-Shot, One-Shot, and Few-Shot prompting across all three models in our study. The same prompt template of different prompting techniques is fol-

lowed for all three models except for Chain-of-Thought prompting. Together, this is a comprehensive collection of common prompting techniques, allowing both a broad overview over the field as well as detailed insights into advantages and disadvantages of single techniques w.r.t. our task of summarizing patient experiences.

Experimental setup

For our experiments, we used the pre-trained model versions of all the models (Flan-T5, GPT-3, GPT-3.5) as shown in Table 2. The table also shows the decoding parameters that were used for all the models across all prompting techniques. For all the models, the summarization length was set to 300 tokens to balance efficiency and completeness while remaining reasonably close to the maximum token length of the ground truth summaries (377 tokens). Following previous work [27], the temperature is set to 0.97 in order to limit the randomness of the generated text in combination with our sampling-based decoding strategy. High temperature values decrease the deterministic behavior of the text-generation model. A value of 0.97 corresponds to a moderate level of randomness, representing a tradeoff between avoiding repetitive and blank summaries and allowing too much randomness and unfaithful generated texts.

Table 2: Pre-trained model versions and decoding parameters of all the models used in the experiment

Models	Pre-trained model versions	Decoding parameters
Flan-T5	flan-t5-base	min_length=150 max_length=300 do_sample=True temperature=0.97 repetition_penalty=1.2
GPT-3	text-davinci-003	temperature=0.97 max_tokens=300 top_p=0.9 presence_penalty=1
GPT-3.5	gpt-3.5-turbo-0301	

In order to test the effect of fine-tuning, the Flan-T5 model was trained using the pre-processed train split of our dataset with the following training parameters: batch_size=1, num_train_epochs=10, learning_rate=2e-5, weight_decay=0.01.

The chosen hyperparameters were set to fine-tune the Flan-T5 model effectively while minimizing the risk of overfitting. A small batch_size=1 helps manage memory limitations and adds implicit regularization. The num_train_epochs=10 allows the model to learn from the data without excessive training, while the low learning_rate=2e-5 ensures stable updates to pre-trained weights, preventing drastic changes that could lead to overfitting. The weight_decay=0.01 acts as regularization by penalizing large weights, further preventing overfitting. Additionally, by using the same decoding parameters for both fine-tuned and non-fine-tuned Flan-T5 models, consistency in evaluation was maintained. To ensure no overfitting, the model's performance was monitored on a validation set, ensuring it didn't memorize the training data.

The fine-tuned Flan-T5 model was used to generate the summaries with the same decoding parameters as shown in Table 2 for the non-fine-tuned Flan-T5 model.

Automatic summaries were generated for all models using the 7 prompting strategies. This was done for the 7 symptoms and 117 treatments, where 1 symptom and 2 treatments were used as examples in the prompt in the few-shot cases, yielding a total of 2,604 summaries (124 summaries by 7 prompting strategies for 3 models). Additionally, for further experiments exploring the influence of fine-tuning, Flan-T5 was trained and validated on 113 summaries (101 summaries for training and 12 summaries for validation) across 8 symptoms and 105 treatments, respectively. The remaining 14 summaries for 14 treatments were used to measure the performance in our automatic evaluation.

Automatic Evaluation

In order to evaluate the generated texts automatically, we rely on state-of-the-art evaluation metrics used in text generation research, in particular the token-based ROUGE and BERTScores. These metrics allow us to compare the similarity of the generated texts to the manually created reference. The token-based ROUGE metrics essentially compute the overlap of phrases between the automatically generated summarization and the reference summarisation, thus estimating the similarity of both summaries. BertScore relies on embeddings and thus compares the summaries at a more semantic level, abstracting from the specific words or phrases that appear.

In the study, the generated summaries from all three models are evaluated using ROUGE, and the rescaled F1-score of BERTScore is calculated using the hidden representation states of the 17th layer of roberta-large without idf-weighting for the generated summaries. For ROUGE, we used three variants: ROUGE-1, ROUGE-2 and ROUGE-L respectively. We employed these three ROUGE variants to thoroughly evaluate generated summaries. Each variant serves a specific purpose: ROUGE-1 measures unigram (single-word) overlap to assess the occurrence of keywords without looking at the context, ROUGE-2 evaluates bigram (pair of consecutive words) overlap to capture context and word associations, and ROUGE-L analyzes the longest common subsequence to evaluate fluency and structural coherence [6]. These metrics collectively provide a nuanced assessment of the generated summary's content, coherence, and adherence to the reference summary.

We present the average score considering all 124 test instances by computing the ROUGE and BERTScores for each reference summary compared with the generated texts. A case study was conducted involving two summaries that were analyzed qualitatively and discussed in more detail, comparing the different generated versions of it in all model settings. In addition to the experiences with LLMs, we also calculate the ROUGE and BERTScores for a baseline using the concatenation of all input texts as a trivial summarization baseline that does not summarize anything.

Results

The evaluation is guided by three research questions:

1. Which model has the best performance compared to the reference summaries?
2. Which prompting techniques are most successful?
3. Does fine-tuning increase its performance in generating patient comment summaries by only using a small amount of training data?

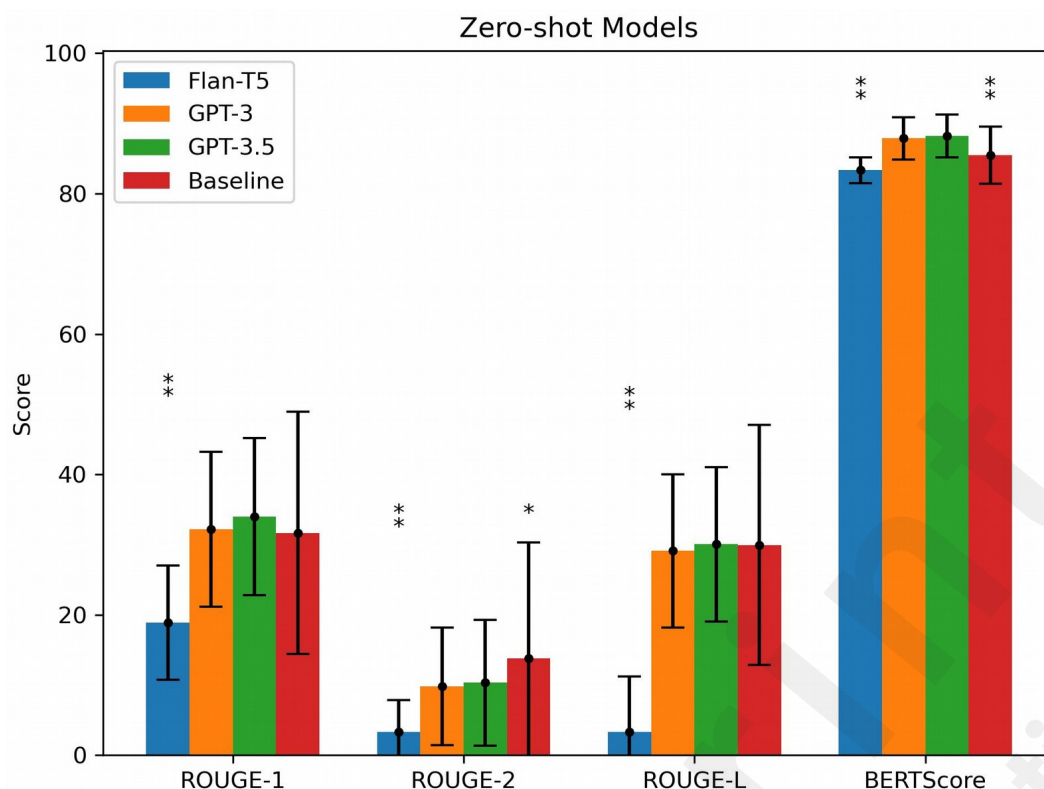


Figure 1: Mean scores per model and metric with corresponding standard deviations as error bars. * for $P < 0.05$, ** for $P < 0.001$. ROUGE Scores and BERTScores obtained for all zero-shot Models without fine-tuning; P values are obtained by comparing to the overall best performing model, i.e., GPT-3.5. The figure shows that GPT-3.5 performs best for all metrics except ROUGE-2, where the baseline performs best.

Regarding the performance of the different models, Figure 1 shows the average ROUGE and BERTScores for 124 summaries generated by using zero-shot prompting on three models. Flan-T5 performs worst for all metrics and with very high significance as well.

In contrast, GPT-3.5 performs better than all other models (incl. the baseline) for all metrics with a margin ranging from 0.4% to 79.8% except ROUGE-2, where the baseline significantly performs best. GPT-3 in all cases performs slightly worse than GPT-3.5 but better than Flan-T5.

The average text length of the input text (and hence, the output of the baseline) is 4,199 words, which is much higher than the average text length of the GPT-3.5 generated summaries, which is 446 words. Therefore, the count of bigrams will be higher in general for the baseline when compared to GPT-3.5 summaries due to its high average text length, which is not suitable for an actual summary. However, in contrast to the baseline, GPT-3.5 generates reader-friendly summaries that are shorter than the original text, having a higher semantic closeness to the reference than the baseline (+2.8 points in BERTScore). A t-test was carried out comparing all other model variants to GPT-3.5 to confirm significance of results. The exact results can be found in the appendix in Table 5.

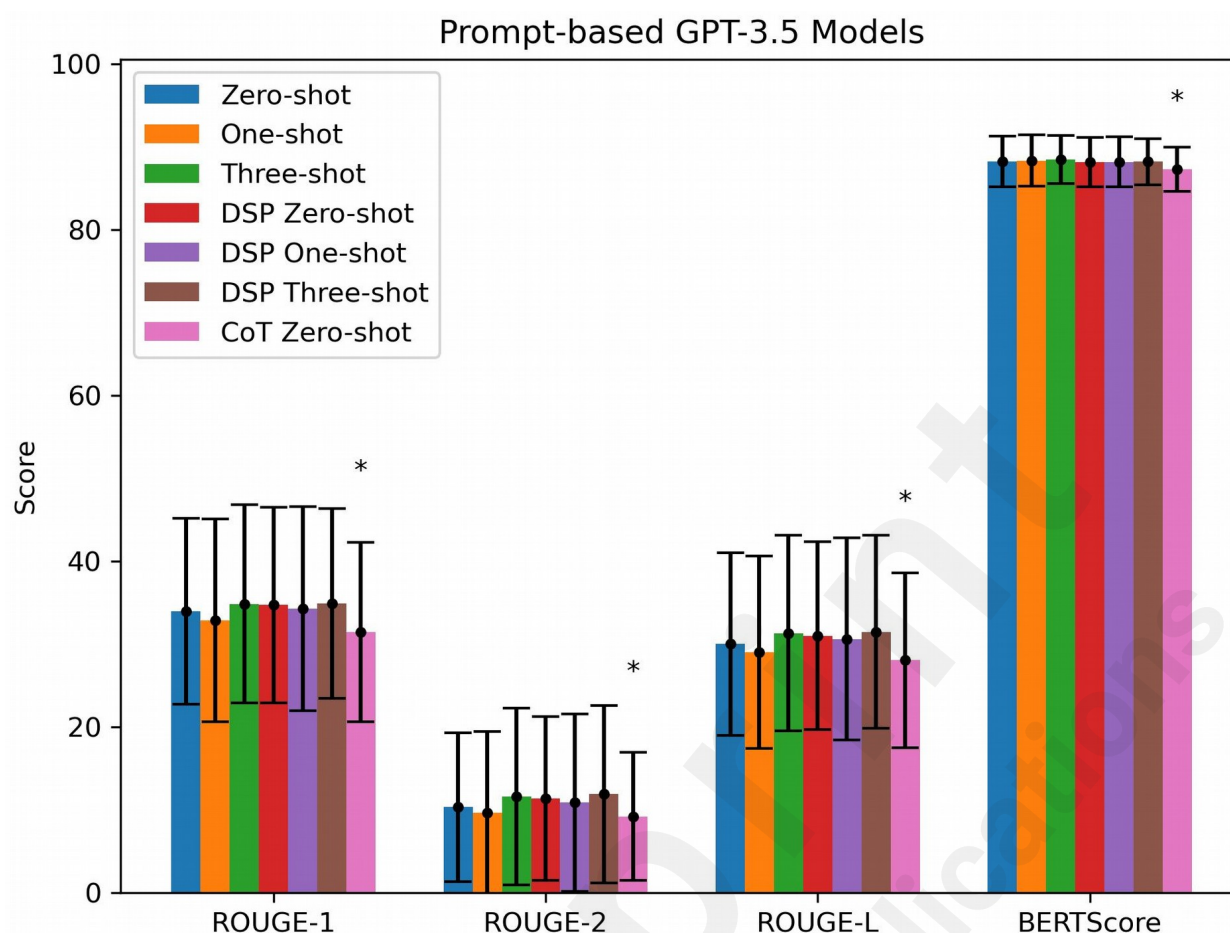


Figure 2: Mean scores per prompting strategy and metric for GPT-3.5 with corresponding standard deviations as error bars. * for $P < 0.05$, ** for $P < 0.001$. ROUGE-1/2/L and BERTScore for prompt-based GPT-3.5 models with P values obtained by comparing all prompt-based GPT-3.5 models to the best performing GPT-3.5 DSP three-shot model. The figure shows that DSP Three-shot performs best. However, none of the results is significant except the worst performance of CoT Zero-shot.

Focusing on GPT-3.5 as the best overall model for zero-shot prompting, we proceed to investigate the impact of different prompting strategies on the summarization performance. The above Figure 2 shows the average ROUGE-1/2/L and BERTScore for 124 summaries generated using the 7 different proposed prompting techniques. Again, the exact results can be found in the appendix in Table 6. Regarding the ROUGE-score, the obtained evaluation scores show that the Directional Stimulus Prompting (DSP) using three examples outperforms other prompting techniques. For BERTScore, three-shot prompting without Directional Stimulus Prompting outperforms all other models. However, the differences between the different prompting techniques regarding their performance are only minor and not significant in most cases, especially for BERTScore. The only significant result is that Chain-of-Thought zero-shot prompting performs worse than all other prompting techniques. The appendix provides tables detailing the results of the seven prompting strategies for Flan-T5 and GPT-3 models.

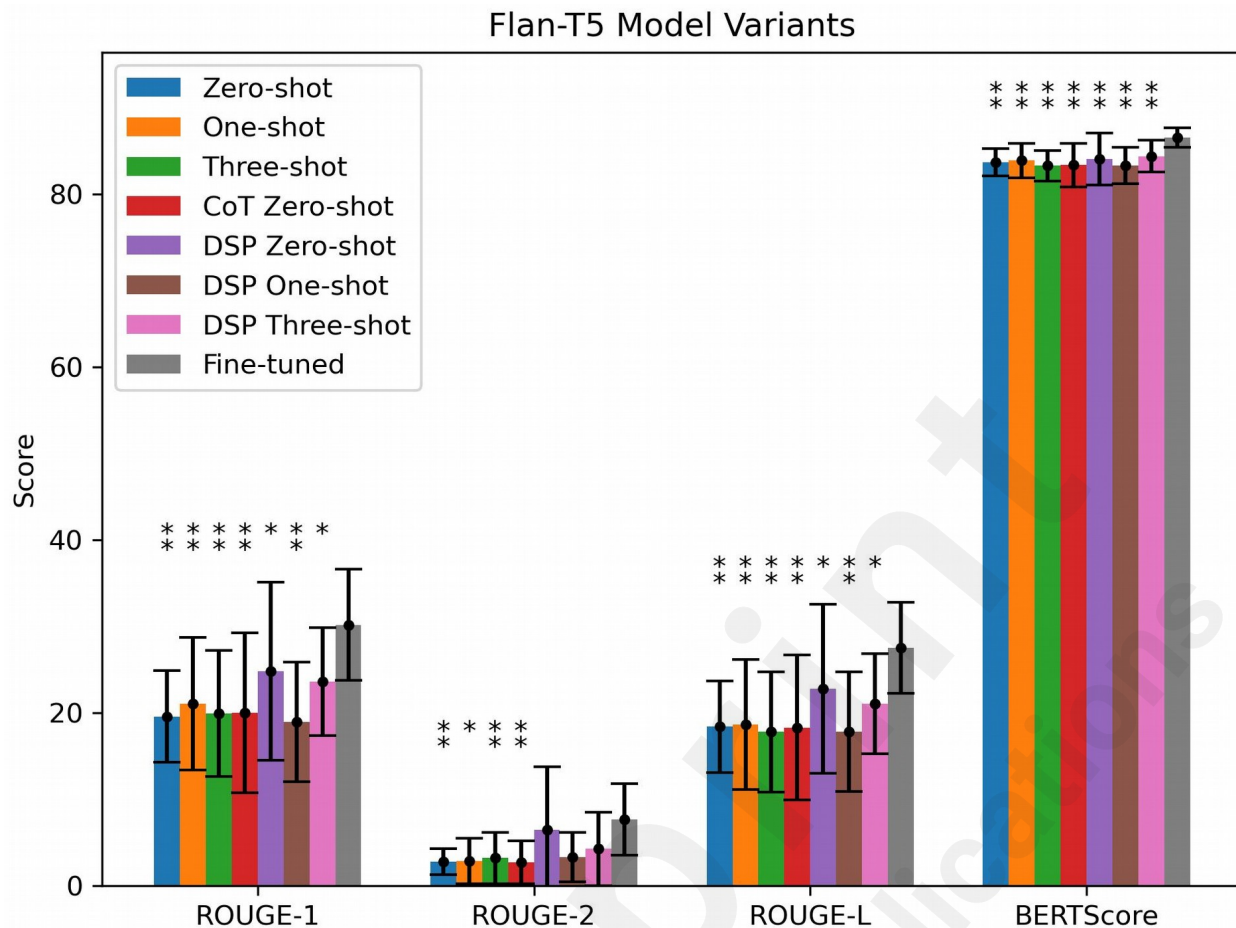


Figure 3: Mean scores per prompting strategy/model variant and metric for Flan-T5 with corresponding standard deviations as error bars. * for $P < 0.05$, ** for $P < 0.001$. ROUGE-1/2/L and BERTScore for Flan-T5 Model variants with P values obtained by comparing all Flan-T5 models to the fine-tuned Flan-T5 model. The figure shows that the fine-tuned model outperforms all non-fine-tuned models no matter what prompting strategy is used.

To explore the effect of fine-tuning under the restriction of having only a small amount of training data, the above Figure 3 shows the average ROUGE-1/2/L and BERTScore obtained for fine-tuned Flan-T5 compared with the prompt-instructed Flan-T5² model for 14 test instances. Again, the exact results can be found in the appendix in Table 7. The fine-tuned Flan-T5 model outperforms the non-fine-tuned Flan-T5 model in all cases, with the non-fine-tuned variants performing worse with normal or even very significant results in almost all cases. Only for the ROUGE-2-scores of the three Directional Stimulus Prompting strategies, the results were not significant. Thus, upon evaluating the obtained scores, it is evident that the fine-tuning - even with only a small amount of training data - is more effective than any prompting strategy for Flan-T5-base.

It is important to note that the fine-tuned Flan-T5 model cannot be directly compared to the prompt-based GPT models; it can only be compared to the prompt-based or non-fine-tuned Flan-T5 models as shown in Figure 3.

² We only consider Flan-T5 for fine-tuning since this GPT-3 and GPT3.5 are closed source, they cannot be fine-tuned locally with a full control over the training process and involved data.

All in all, among the tested models, GPT-3.5 performs best. The best-performing prompting strategy is DSP Three-shot for the ROUGE scores and Three-shot for BERTScore. For Flan-T5, the fine-tuned version performs better than all prompting strategies on non-fine-tuned Flan-T5 models.

Case Study

In order to discuss the performance of the approach in a more qualitative manner, we analyze the different summaries provided by the different models in comparison to the reference summary for one treatment (Carboplatin) and one symptom/phenotype (Neoplasm). We selected examples where the ROUGE scores were highest. While it might seem unfair to select the topics with the best results, the case study is not meant as a proper evaluation, but only to provide examples of how the summaries differ across methods.

In Table 3 we see the different summaries on the patient experience with Carboplatin, once the reference summary and the summaries provided by GPT-3.5 Zero-shot, GPT-3.5 Three-shot and GPT-3.5 DSP Three-shot, respectively. The **reference summary** emphasizes that Carboplatin, being a new treatment and good option to treat TNBC (Triple Negative Breast Cancer), has **good study results** and **better tolerability, and is getting increasingly popular**. It emphasizes the **side effects** such as **hair loss** and **heart damage**. It mentions that side effects might increase in combination with other drugs such as Taxol and Gemcibine.

The **GPT-3.5 Zero-shot summary** highlights the **efficacy** (although not good study performance explicitly), **side effects** (hair loss but in addition hearing loss and tinnitus) and **popularity of the treatment**. In addition, it also highlights **better tolerability** of carbo taxol (!) in comparison to carbo alone. It provides further details on dosage and scheduling and even goes as far as talking about the mechanism of action.

The **GPT-3.5 Three-shot Summary** highlights that the Carboplatin has had good study results as well as the side effects, mentioning hearing loss and tinnitus. The summary misses heart damage, but includes an additional side effect not mentioned in the reference summary, i.e. tinnitus. The summary goes further than the human-generated summary in terms of discussing regime and scheduling in addition to discussing outcomes in combination with other drugs.

The **GPT-3.5 DSP Three-shot Summary** emphasizes the side effects (again hearing loss and tinnitus) as well as better tolerability. It highlights that Carboplatin has shown promising results in clinical trials. Beyond that, it provides more detail on how patients deal with the side effect by using other treatments.

Overall, the three summaries cover the most important points highlighted by the human reference summary, providing some additional details (dosage, mechanisms of action, other side effects, combination with other drugs or treatments to reduce side effects, etc.).

Table 3: Reference summary, GPT-3.5 Zero-shot summary, GPT-3.5 Three-shot summary and GPT-3.5 DSP Three-shot summary about Carboplatin. The table shows that all summaries mention the most important points.

Reference Summary about Carboplatin: *"Carboplatin is a more recent treatment, deemed by patients as a good option to treat TNBC, having encouraging results in studies and a better tolerability. The treatment is more novel, yet becoming more popular. But it also causes side effects such as hair loss, heart damage. When combined with other drugs such as Taxol and Gemcibine, the side effects increase."*

GPT-3.5 Zero-shot Summary: *"The patient comment covers a variety of topics related to Carboplatin, a chemotherapy drug used to treat Triple Negative Breast Cancer (TNBC). The comment discusses the efficacy of Carboplatin in treating TNBC even without HER2 positivity, potential side effects such as hearing loss and tinnitus, and its popularity as a treatment option. The comment also touches on the dosing schedule and suggests that weekly carbo taxol is easier to tolerate than every 2 weeks IV carbo. Some comments mention hair loss and other side effects associated with Carboplatin. The comment also highlights the role of calcium carbonate in reducing the absorption of other medicines from the gut."*

GPT-3.5 Three-shot Summary: *"Carboplatin is a chemotherapy drug used to treat Triple Negative Breast Cancer (TNBC), even if the patient does not have Her 2 positivity. Patients report that carboplatin has been shown to have good results in trials and may be a better option than other treatments such as AC. However, some patients also report potential side effects of carboplatin, such as hearing loss and tinnitus. Carboplatin can be given as a weekly regime or every two weeks, but the latter may cause a decrease in hemoglobin levels. Despite this, carboplatin is still seen as an effective treatment option for high-grade TNBC. There is also ongoing research into the use of carboplatin with other drugs to improve treatment outcomes."*

GPT-3.5 DSP Three-shot Summary: *"Carboplatin is a commonly used chemotherapy drug for treating TNBC, although it can cause side effects such as hearing loss and tinnitus. Patients report mixed experiences with the drug, with some finding it tolerable and others experiencing difficulty with its side effects. Carboplatin has shown promising results in clinical trials and is being studied in combination with other agents such as leronlimab. It is not clear why some patients are prescribed carboplatin over other chemotherapy drugs, but it may be preferred for high-grade TNBC or in combination with radiotherapy. Patients also report using bicarbonate and calcium carbonate to manage side effects and improve blood pH levels during treatment."*

In Table 4, we observe different summaries on the patient experience with Neoplasm, including the reference summary and the summaries provided by GPT-3.5 Zero-shot, GPT-3.5 Three-shot, and GPT-3.5 DSP Three-shot. The reference summary highlights patients experiences with neoplasm, **focusing on tumor size, aggressiveness, and growth rate**. It discusses **tumor structure**, including **receptors and mutations**, and **varied treatment responses**. Patients reported problems like **fluid retention, lymphedema, tissue damage, severe pain**, and complications from spinal spread, such as **vertebral fractures and bone marrow impairment**. Treatment outcomes varied, with some tumors decreasing, staying stable, or growing despite interventions.

The *GPT-3.5 Zero-shot summary* discusses patient experiences with neoplasms, covering **tumor characteristics, treatment options, and responses**. It distinguishes between benign and malignant tumors and mentions the potential for immunotherapy. However, it **lacks details on specific patient-reported problems and complications** highlighted in the reference summary, such as fluid retention and bone marrow impairment due to spinal spread.

The *GPT-3.5 Three-shot summary* echoes the reference summary's emphasis on patients' experiences with **tumor characteristics, including size, grade, and aggressiveness**. It acknowledges the importance of post-surgical treatments like chemotherapy and radiation in preventing cancer recurrence. It also discusses tumor markers and lymph node involvement. However, it provides a

more concise overview, **focusing on patients' concerns about tumor growth and aggressiveness**, and the importance of treatment in improving survival rates.

The *GPT-3.5 DSP Three-shot summary* delves into patients' **tumor descriptions, covering size, stage, grade, and lymph node involvement**. It **acknowledges larger, aggressive tumors necessitating post-surgical treatments** like chemotherapy or radiotherapy. Additionally, it recognizes benign tumors and the significance of tumor markers such as ER, PR, and HER2 in treatment decisions. The summary also discusses tumor flare post-biopsy and size fluctuations over time, stressing the uniqueness of each patient's tumor and the need for personalized treatment plans, echoing the reference summary's focus on varied tumor responses and patient challenges.

Overall, the three summaries cover the key points highlighted in the reference summary, while providing additional details such as genetic mutations, treatment responses, and the importance of individualized treatment plans.

Table 4: Reference summary, GPT-3.5 Zero-shot summary, GPT-3.5 Three-shot summary and GPT-3.5 DSP Three-shot summary about Neoplasm. The table shows that again all important details are mentioned in the summaries, such as genetic mutations and treatment responses.

Reference Summary about Neoplasm: *"Regarding neoplasm, patients often describe the size of their tumors, how aggressive they are and how fast they grow. Furthermore, they also describe their structure, which receptors, mutations they have and how fast or slow they respond to treatments. There were many varied responses in these regards, e.g. many small but also many large treatments, responsive and non-responsive treatments and different types of tumors. Neoplasms caused an array of problems for patients such as fluid retention, lymphedema, acidity, tissue dama, a concave chest after removal of the tumor and chest muscle, inflammation, compression of nerves, severe pain and an immunosuppressive environment. For some patients where it spread to their spine, it impaired their bone marrow, caused complications, broke vertebrae and aggravated nodes. Some neoplasms decreased with treatment, but others either decreased and then came back, stayed stable or even grew despite treatment."*

GPT-3.5 Zero-shot Summary: *"The patient comments mention various sizes and stages of tumors, as well as the presence or absence of lymph node involvement and markers such as HER2. Some tumors are benign while others are malignant, and treatment options include surgery, chemotherapy, and radiotherapy. Response to treatment varies, with some tumors showing complete response and others being slow-growing. In some cases, genetic mutations may be acquired under the selective pressure of therapy, and hot tumors with inflammation may respond to immunotherapy. Some tumors contain a mix of different cell types, and there may be concordance or discordance between cell-free DNA and tissue samples. Overall, the comments suggest that cancer diagnosis and treatment can be complex and individualized, and it is important to trust medical professionals while seeking support from others who have gone through similar experiences."*

GPT-3.5 Three-shot Summary: *"Patients share their experiences with tumors, including their size, grade, and aggressiveness. They also mention the importance of post-surgical treatments, such as chemotherapy and radiation, in preventing cancer recurrence and improving survival rates. Patients highlight the difference between benign and malignant tumors and share their tumor markers and lymph node involvement. Some patients express concern about the growth and aggressiveness of their tumors, while others*

report slow-growing tumors."

GPT-3.5 DSP Three-shot Summary: *"Patients describe their tumors in terms of size, stage, grade, and involvement of lymph nodes. Some patients report having larger and more aggressive tumors, which require post-surgical treatments such as chemo or radiotherapy to manage them. However, some tumors are benign and do not necessarily indicate cancer. Patients may also have a combination of different tumor markers such as ER, PR, and HER2, which may affect the aggressiveness of the cancer and the recommended treatment options. Tumor flare is a real phenomenon that can occur after biopsy, and tumor size can fluctuate over time. Some patients also report elevated tumor markers in their blood. Ultimately, each patient's tumor is unique and requires an individualized treatment plan."*

Discussion

Our evaluations of different models on the task of generating summaries about the patient experience has revealed that GPT-3.5 is the best-performing model. This is no surprise since it is the model with the most parameters (~375B) in the set of models examined. GPT-3 has 175B parameters, and the smallest model is Flan-T5 with 0.2B parameters in the version we used. As the training data sources for the GPT models are not made public, we can not attribute the differences in performance to differences in the training data.

Taking GPT-3.5 as the best performing model (see research question 1), we have examined in detail the impact of different prompting strategies. For the three ROUGE metrics, DSP Three-shot has emerged as the best strategy, while Few-shot with three examples received the best scores according to BERTScore (see research question 2). In any case, the results clearly show that providing at least three examples helps the model to generate more tailored summaries. DSP Three-shot benefits particularly from the fact that key relevant topics are provided as guidance to the model, which helps to generate better summaries.

Our fine-tuning experiments with Flan-T5 have shown that fine-tuning can have a positive effect in principle (see research question 3). However, we assume that the amount of data was still too low. The fine-tuned Flan-T5 setting is therefore not really comparable to the other settings which have not been fine-tuned with a comparable number of examples. In particular, the results of Flan-T5 are, irrespective of fine-tuning, quite lower compared to the other models. Yet, the results show the potential of fine-tuning within a particular model class.

Our case study using GPT-3.5 as best performing model with the best prompting strategies has shown that the summaries produced by GPT-3.5 are reasonable, covering most of the aspects of the manually-generated summaries while adding some additional details or background. Our results demonstrated that such models can be used to speed up qualitative research by being able to summarize large amounts of texts automatically. In addition, we confirmed a positive correlation between the quality of summaries and models' size on the one hand and fine-tuning on the other hand. In March 2023, OpenAI released GPT-4 as well as the possibility to fine-tune models using a fee-based API. These new developments are promising and should be investigated further, yet they were not available at the time the experiments described in this paper were carried out.

Taken together, our results suggest that state-of-the-art pre-trained language models can be a valuable tool to provide qualitative insights about patient experience to better understand unmet needs, patient priorities, how a disease impacts daily functioning and Quality of Life in order to inform processes aimed at improving healthcare delivery and ensure that drug development focuses more on the actual priorities and unmet needs of patients. In particular, the insights would allow to focus drug development and clinical development on outcomes that are of relevant to patients and have the potential to significantly improve their Quality of Life. The insights provided by our method could be used to develop a more detailed understanding of patient journeys as a basis to improve patient care by focusing on the actual problems and unmet needs that patients face in different phases of their journey. The insights could further help in developing support programs that provide relevant information to patients when they need it, to empower them to make better decisions. Finally, such qualitative insights could also inform the design and conceptualization of digital apps for patients.

As a notable limitation of the present study, our manually-created test set for quantitative evaluation of the different approaches is of limited sample size. Although we present an experiment with a fine-tuning approach, splitting the small dataset into a train-, development- and test-split, we notice that each split is too small to derive representative results for each setting. Having a larger dataset (at least by one order of magnitude) would enable more experiments with respect to fine-tuning. While the sample size was small, it is a remarkable result that fine-tuning had a positive impact on the task. A larger dataset would possibly have yielded even stronger results. However, since the focus of this paper is to experiment with (unsupervised) prompting approaches, we keep experiments with a larger dataset for future work. A further important limitation is the fact that the reference summaries have been provided by only one annotator, which might bias the results to the perspective of this single annotator. The huge effort involved in creating these summaries has not made it possible to have multiple subjects providing the summaries. A larger dataset could be obtained via crowdsourcing platforms such as CrowdFlower or Prolific; a drawback of these platforms is however that we can not control for the (domain) expertise of labelers reliably.

Summaries can not be directly compared using standard annotation metrics such as Cohen's Kappa, but could have been compared in terms of ROUGE to each other to have an upper baseline for the scores.

Another limitation arises from the nature of our applied automatic metrics. The average text length of the output of the baseline approach (mirroring its input) is substantially higher (contradicting the idea of a summary that should be concise) than that of the GPT-3.5 generated summaries. However, in terms of our automatic metrics, this discrepancy in text length impacts the count of bigrams, favoring the baseline compared to GPT-3.5 zero-shot. The automatic evaluation metric we used, for example, ROUGE-2, does not consider the text length, which may bias the results towards the longer text. Thus, while the baseline appears to perform better according to ROUGE-2, it is important to interpret these results cautiously, considering the difference in text length.

We have compared the different LLMs and the different prompting strategies on our own dataset consisting of the 127 reference summaries. The dataset might not be completely representative for all sorts of summarization tasks, possibly limiting the generalizability of results to other related tasks. The comparison of the different methods and the results obtained are specific to the particular summarization task considered. It is an open question if the results generalize to other summarization tasks. Our focus has been however on determining whether LLMs can reliably and accurately summarize the experience of patients as shown on online posts. A question for further research is to

better understand why the performance of models differs and which underlying factors are responsible for the differences. Understanding whether the differences in ROUGE and other scores represent semantically meaningful differences is also an important question for future work. Indeed, future work might strive for a more semantic assessment of the generated summaries by manual coders that could rate the generated text in terms of variables such as i) clarity, ii) comprehensiveness, iii) completeness, iv) readability, etc.

Beyond this, since the time our experiments have been carried out, newer language models have been released, in particular also GPT-5 and GPTo, which could also be tested. Other open source models such as Llama could also be investigated.

To the best of our knowledge, our dataset is the only dataset consisting of manually created summaries for patient experience. We can thus not evaluate our method on existing datasets. To allow researchers to compare their methods on our dataset, we will provide the dataset to anyone interested upon request.

Beyond improving the methodology and data basis and models used, future research should show that the insights derived via summarization methods can indeed impact patient insight generation and positively impact drug development, optimization of patient journeys etc. This question is clearly out of scope of this article, but could be pursued in follow-up work. While we have focused only on posts from breast cancer patients, we see no reason why the approach could not generalize to other patient populations. This however needs to be validated in future work.

In carrying out this study, we have ensured that posts are automatically anonymized to ensure that the identity of patients is protected. Regarding data protection, our methodology is compliant with the General Data Protection Regulation (GDPR). According to the GDPR, there is a trade-off between the legitimate interests of data processors to process data that needs to be balanced against the data protection rights of data subjects. In this particular case, there is a legitimate interest by society to rely on public data shared by patients as this data can contribute to improving the standard of care as well as the development of drugs that are relevant and address outcomes of relevance to patients. Further, given that patients have made their data “overtly and explicitly public” without any additional protection, we can assume that the sensitivity of shared data is not that critical, so that exploiting it for the legitimate purpose of improving health care services is acceptable. The GDPR foresees that consent has to be obtained from data subjects to process their data. Given that data was anonymized, patients can not be directly contacted to obtain consent. In this respect, we process data making use of the exemption mentioned in the GDPR for processing data without explicit consent in case that obtaining consent is not possible “with a reasonable effort”. Obtaining consent would in our case not represent a reasonable effort as each patient would have to be contacted individually, some patients might not even be active anymore on the fora or sites in question. Contacting specific users would require registering on the platform in question, which in many cases would violate the terms and conditions of the platform as only patients are allowed to register. On many fora, direct messages can not be sent to users, so that they could not be directly contacted. In general, it has been a matter of debate whether observational studies can be ethical without patient consent, and which conditions would need to be met for the absence of patient consent to be ethical. The UK’s Economic and Social Research Council (ESRC) for instance considers ethical the forgoing of patient consent as long as a study is not “undertaken lightly or routinely”. It is only justified if important issues are being addressed and if matters of social significance which cannot be uncovered in other ways are likely to be discovered” (102). A broader review of ethical recommendations [47] suggests that SML for health research can be ethical without informed consent in case that

anonymization is ensured, harms are minimized and there is a clear public benefit resulting from the analysis. As argued above all three criteria are fulfilled for our study.

Related Work

We briefly discuss related work on using LLMs in healthcare in general, deep diving into work on using natural language processing for analyzing patient experiences in social media. We discuss application of summarization in healthcare as well.

LLMs in Healthcare: Very recently, large language models have been considered for different tasks and use cases in healthcare. In a recent overview paper [39] Clusmann et al. have described the main application areas in which there is the potential to apply LLMs in healthcare: research, patient care and education. The use of LLMs to summarize patient discussions falls into the area of patient care as it has the potential to discover unmet needs of patients that can contribute to improve drug development and patient care.

Beyond text summarization tasks, LLMs have been also investigated on use cases related to drug discovery [41] and have been tested on tasks such as helping to discover new targets as well as predicting characteristics of drugs including molecular pharmacodynamics (PD), pharmacokinetics (PK), and toxicity [42]. They have also been shown to be able to support the discovery of drug-drug interactions [43]. A recent review has even highlighted the opportunities, limitations and barriers for the use of LLMs in medical diagnosis [44].

Social Media Listening: Social media listening (SML) has been established as a mature methodology to extract key insights about the patient experience as a way to inform drug development and improvement of patient care. A recent review has analyzed 63 publications that performed SML [28]. Out of 63 publications analyzed, 38 carried out some computer assisted analysis. The used methods cover a wide bandwidth of NLP methods to detect topics and concepts within textual data. While some published research used keywords and dictionaries to identify topics in some cases (e.g. [29]), other methods comprise Binary Matrix Factorization, Bayesian Estimation, Latent Dirichlet Allocation [31]. Several publications state that custom trained NLP algorithms were used (e.g. [30], [32]) and also neural networks were reportedly used to identify certain topics or expressions [33]. 25 out of 63 publications performed manual analysis, relying on the manual coding of small data samples. The methods that analyze social media posts automatically by using NLP methods typically extract certain pre-defined aspects of the patient experience including key symptoms, their severity, experiences with treatments etc. While they support the scalable analysis of large amounts of texts, they are typically neither suited to analyze open topics nor provide detailed qualitative insights. Manual analysis methods that rely on the development of a specific coding can be used to answer very specific research questions from a qualitative perspective. However, they are not designed to scale and are necessarily limited to analyzing a smaller sample. The method proposed in this article can be understood as a sweet spot between these two methodological ends of the spectrum. The automatic summary of patient experience conversations allows to scale to larger amounts of data and text while allowing to succinctly summarize the data to support qualitative understanding.

Automatic Text Summarization: Summarizing texts is an important field in the area of natural language processing, aiming to condense a verbose text into a shorter version while preserving the essence of its meaning. While the former research started with extractive summarization, i.e. selecting sentences that are concatenated in order to yield a summary, with the rise of large language models, the research has shifted towards abstractive summarization, in which a model generates a text instead of copying text parts [26], resulting in a fluent text that helps the reader to get an overview with minimal time investment. To this end, the input text is encoded into an internal representation which is enrolled by a decoder into natural language. By massive pre-training imitating summarization tasks [20], fine-tuning models towards summarization [24], or tailoring them towards instructions asking for summarizations [19], this task becomes a deeply explored field in natural language processing.

In the domain of healthcare, automatic summarization of patient data has emerged as a vital area of research. Here we explore the recent advancements in summarization techniques of patient data and other types of data, highlighting the models and methodologies employed in alignment with our research.

It has been previously investigated how LLMs can support tasks related to medical documentation and information summarization. LLMs have been particularly evaluated on the summarization of radiology reports, patient questions, progress notes, and doctor-patient dialogue [40]. The objectives of the study by Van Veen et al. were similar to ours in that the goals was to evaluate the effectiveness of Large Language Models on the task of summarizing clinical text. In contrast to our focus on patient experience reports from social media or patient fora, their work focuses on other types of data (radiology reports, patient questions, progress notes and dialogues). They used similar metrics for the evaluation of the automatically generated summaries compared to manual summaries, i.e. ROUGE_L and BERTScore, in addition to BLUE. While they also examined similar models, Flan-T5 as well as GPT (using version 4 instead of 3.5), they did not examine an exhaustive set of prompting strategies as we do. They also concluded that GPT had the best performance, generating summaries that were more complete on average than human summaries. These results are in line with our results in this study.

A recent study testing the ability to summarize medical evidence using LLMs has shown clear limits of these technologies, as factual inconsistencies produced by LLMs are critical in some applications such as the one of summarizing medical evidence, which requires increased care and accuracy as such results might inform the decision making of many [46]. In their study, Tang et al. [46] focused on the task of medical evidence summarization referring to the process of extracting and synthesizing key information from a large number of medical research studies and clinical trials into a concise and comprehensive summary. They also used automatic evaluation metrics including ROUGE-L, BLUE and METEOR to evaluate the automatically generated texts with respect to expert-created Cochrane Reviews. They explored models from the GPT family only (GPT3.5 and ChatGPT), but did not analyze the impact of different prompting strategies, relying only on Zero-Shot prompting. In terms of ROUGE-L scores, GPT-3.5 performed better than ChatGPT, although this was not the case for other metrics.

A study by Koyena Pal et al. [14] investigated the effectiveness of various neural network architectures in generating hospital discharge summaries from Electronic Health Records (EHRs) within the healthcare domain. They applied different models (BART, Longformer, T5, FLAN-T5) and examined the impact of fine-tuning on the results, relying on similar automatic evaluation metrics as we do (ROUGE-1/2L). In contrast to our work, they did not use any instruction-following models from the GPT family, so that they could not investigate the impact of different prompting strategies. In line with our results, their study shows that fine-tuning can consistently improve results.

Another research study by Varun Nair et al. [15] proposes a multi-stage approach that leverages few-shot prompting techniques to generate summaries of patient-provider dialogues within the healthcare domain. They relied on GPT-3 as a model for their multi-stage approach and validated their method with respect to 100 clinical encounters. They examined in particular the impact of the number of examples provided in the prompt, varying between 1, 3 and 5 examples.

The effectiveness of pre-trained models for text summarization is further supported by Betul Ay et al. [16]. Their work demonstrates the successful application of the T5 model, pre-trained on news data, for abstractive text summarization of Turkish text documents. In contrast to our work, they did not focus on summarization of health-related texts and only relied on one sequence to sequence model, T5. In contrast, we have also examined auto-regressive models from the GPT family and investigated the impact of different prompting strategies.

The study by Yiming Wang et al. [4] introduces a novel element-aware summarization approach using GPT-3 and Chain-of-Thought prompting for abstractive text summarization. While they worked on a completely different domain (news texts rather than health-related texts), they follow a similar methodology, relying on the same evaluation metrics (ROUGE-1/2/L as well as BERT Score), also examining the impact of fine-tuning BART and T5 compared to using GPT-3 with zero-shot prompting. Their results were mixed, depending on the dataset. As a general trend, they showed that the larger models such as GPT-3 perform better than the smaller (but fine-tuned) models. This study is the only one in addition to ours investigating more advanced prompting strategies such as Chain-of-Thought.

Li et al. [5] have introduced Directional Stimulus Prompting (DSP), which we have used in our experiments, as a method to guide large language models for text summarization tasks. They have evaluated their proposed method on the task of summarizing news and also evaluated the suitability of the generated summaries using BLEU, ROUGE, Meteor and BERTScore. In terms of models they also focus on instruction-following models such as GPT-3, InstructGPT and ChatGPT. They have shown that all the models can benefit from the use of Directional Stimulus, which is in line with our results.

Taken together, we see that none of the existing works has actually focused on summarizing patient reports from social media or fora. Most studies have also used automated metrics such as ROUGE, BLUE, METEOR and BertScore as we do, in some cases complementing this with a manual evaluation by experts. In terms of the models investigated, our study is similar to others in the sense that in most studies encoder/decoder-based sequence-to-sequence models have been investigated, including in particular T5 and FLAN-T5, as well as auto-regressive pre-trained decoder-only models from the GPT family (GPT3, GPT-3.5, GPT4 and ChatGPT). In line with our results, most studies

have shown the superiority of the GPT family models. Our work is the only one that has investigated the impact of very different prompting strategies, showing in particular the impact of using Direct Response Stimulus in combination with few-shot training. In terms of comparison, it is not possible to directly compare our results to other studies as none of the above mentioned studies have focused on patient reports.

Conclusion

Exchanging information with peers on social media sites, support fora or social networking as well as receiving emotional support from these is important for patients in coping with their condition and empowering. Yet, the information they share on public sites is unstructured and vast, not accessible for those interested in using the information to improve health care. Methodologies for generating qualitative information at large scale from these dataset have so far been missing.

Towards closing this gap, this article shows that it is possible to accurately and reliably summarize the content shared by patients in social media sites, social networking sites, discussion forums and support groups automatically. Our article has explored the feasibility of automatically summarizing only patient contents using generative language models combined with different strategies of prompting. Our comparative study highlights the effectiveness of different advanced prompting techniques used with generative language models in summarizing online patient comments. Notably, few-shot models, particularly the Three-shot variants, exhibit superior results compared to their Zero-shot and One-shot counterparts. Within the Few-shot category, GPT-3.5 stands out with higher evaluation scores compared to other models. In the context of Directional Stimulus Prompting (DSP) models, DSP Few-shot (three-shot) models, specifically GPT-3.5, deliver the most favorable results. Upon comprehensive comparison, Directional Stimulus Prompting-based Three-shot models emerge as the most efficient prompt-based model for summarization tasks, especially in the context of patient comment summarization. Although our experiments fine-tuning the smaller model Flan-T5 increases its performance by 50% in terms of the longest common subsequence to the reference, our advanced prompting-techniques using larger models as GPT-3.5 are superior in the explored domain.

Our results open up the possibility for using the qualitative insights generated at a larger scale to inform processes of drug development, design of clinical trials, development of patient journeys, public policy making and informing the development of digital health apps. Information about unmet needs, symptoms and their impact on Quality of Life of patients, disease burden, experiences with existing treatments all provide valuable insights to inform the above mentioned processes and ultimately improve the state of care for patients.

Acknowledgements

None

Abbreviations

OSP : 0-Shot Prompting

1SP : 1-Shot Prompting

BART : Bidirectional and Auto-Regressive Transformers

BERT : Bidirectional Encoder Representations from Transformer

CoT : Chain Of Thought Prompting

DSP : Directional Stimulus Prompting

FSP : Few-Shot Prompting

GPT : Generative Pretrained Transformer

LLMs : Large Language Models

NLP : Natural Language Processing

ROUGE : Recall-Oriented Understudy for Gisting Evaluation

RoBERTa : Robustly Optimized BERT Pretraining Approach

T5 : Text-to-Text Transfer Transformer

References

1. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S. and Webson, A., 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70), pp.1-53.
1. Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training.
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp.1877-1901.
3. Wang, Y., Zhang, Z. and Wang, R., 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.
4. Li, Z., Peng, B., He, P., Galley, M., Gao, J. and Yan, X., 2024. Guiding large language models via directional stimulus prompting. *Advances in Neural Information Processing Systems*, 36.
1. Chin-Yew, L., 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, 2004.
2. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y., 2019. BERTScore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
3. Grootendorst, M., 2020. KeyBERT: Minimal keyword extraction with BERT. Zenodo. doi: 10.5281/zenodo.4461265.
5. Schopf, T., Klimek, S. and Matthes, F., 2022. Patternrank: Leveraging pretrained language models and part of speech for unsupervised keyphrase extraction. *arXiv preprint arXiv:2210.05245*.
6. Zhu, P., Shen, J. and Xu, M., 2020. Patients' willingness to share information in online patient communities: Questionnaire study. *Journal of medical Internet research*, 22(4), p.e16546.
7. Chen, J. and Wang, Y., 2021. Social media use for health purposes: systematic review. *Journal of medical Internet research*, 23(5), p.e17917.

8. De Martino, I., D'Apolito, R., McLawhorn, A.S., Fehring, K.A., Sculco, P.K. and Gasparini, G., 2017. Social media for patients: benefits and drawbacks. *Current reviews in musculoskeletal medicine*, 10, pp.141-145.
9. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M. and Le, Q.V., 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
10. Pal, K., Bahrainian, S.A., Mercurio, L. and Eickhoff, C., 2023. Neural Summarization of Electronic Health Records. *arXiv preprint arXiv:2305.15222*.
11. Nair, V., Schumacher, E. and Kannan, A., 2023. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. *arXiv preprint arXiv:2305.05982*.
12. Ay, B., Ertam, F., Fidan, G. and Aydin, G., 2023. Turkish abstractive text document summarization using text to text transfer transformer. *Alexandria Engineering Journal*, 68, pp.1-13.
13. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
14. Merchant, A., Rahimtoroghi, E., Pavlick, E. and Tenney, I., 2020. What happens to BERT embeddings during fine-tuning?. *arXiv preprint arXiv:2004.14448*.
2. OpenAI, 2022. ChatGPT: Optimizing Language Models for Dialogue. OpenAI.com, 30 November.
15. Goodwin, T.R., Savery, M.E. and Demner-Fushman, D., 2020, December. Flight of the PE-GASUS? Comparing transformers on few-shot and zero-shot multi-document abstractive summarization. In *Proceedings of COLING. International Conference on Computational Linguistics* (Vol. 2020, p. 5640). NIH Public Access.
16. Rothe, S., Maynez, J. and Narayan, S., 2021, November. A thorough evaluation of task-specific pretraining for summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 140-145).
17. Hosseini-Asl, E. and Liu, W., Salesforce com Inc, 2023. *Generative language model for few-shot aspect-based sentiment analysis*. U.S. Patent 11,853,706.
18. Salnikov, M., Le, H., Rajput, P., Nikishina, I., Braslavski, P., Malykh, V. and Panchenko, A., 2023. Large Language Models Meet Knowledge Graphs to Answer Factoid Questions. *arXiv preprint arXiv:2310.02166*.
19. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), pp.1-67.
20. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N. and Mian, A., 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
21. Liu, Y. and Lapata, M., 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.
22. Wiher, G., Meister, C. and Cotterell, R., 2022. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10, pp.997-1012.
23. Cimiano, P., Collins, B., De Vuono, M.C., Escudier, T., Gottowik, J., Hartung, M., Leddin, M., Neupane, B., Rodriguez-Esteban, R., Schmidt, A.L. and Starke-Knäusel, C., 2024. Patient listening on social media for patient-focused drug development: a synthesis of considerations from patients, industry and regulators. *Frontiers in Medicine*, 11, p.1274688.

24. Lognos, B., Carbonnel, F., Launay, I.B., Bringay, S., Guerdoux-Ninot, E., Mollevi, C., Senesse, P. and Ninot, G., 2019. Complementary and alternative medicine in patients with breast cancer: exploratory study of social network forum data. *JMIR cancer*, 5(2), p.e12536.
25. Picone, M., Inoue, S., DeFelice, C., Naujokas, M.F., Sinrod, J., Cruz, V.A., Stapleton, J., Sinrod, E., Diebel, S.E. and Wassman, E.R., 2020. Social listening as a rapid approach to collecting and analyzing COVID-19 symptoms and disease natural histories reported by large numbers of individuals. *Population Health Management*, 23(5), pp.350-360.
26. Quartey, G., Ley-Acosta, S., Pierce, C., Nguyen, A. and Ertle, G., 2018. Social media listening for patient risk characterization. *Value in Health*, 21, p.S115.
27. Spies, E., Andreu, T., Koelling, J., Hartung, M., Kamudoni, P. and Park, J., 2022, December. Retrospective Social Listening Study of Patients Living with Systemic Lupus Erythematosus (SLE): Understanding the Patient Experience. In *Value in Health* (Vol. 25, No. 12, pp. S438-S438). STE 800, 230 PARK AVE, NEW YORK, NY 10169 USA: ELSEVIER SCIENCE INC.
28. Weissenbacher, D., Ge, S., Klein, A., O'Connor, K., Gross, R., Hennessy, S. and Gonzalez-Hernandez, G., 2021. Active neural networks to detect mentions of changes to medication treatment in social media. *Journal of the American Medical Informatics Association*, 28(12), pp.2551-2561.
29. Smailhodzic, E., Hooijsma, W., Boonstra, A. and Langley, D.J., 2016. Social media use in healthcare: A systematic review of effects on patients and on their relationship with health-care professionals. *BMC health services research*, 16, pp.1-14.
30. Suler, J., 2004. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3), pp.321-326.
31. Ma, X., Hancock, J. and Naaman, M., 2016, May. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 3857-3869).
32. Seltzer, E., Goldshear, J., Guntuku, S.C., Grande, D., Asch, D.A., Klinger, E.V. and Merchant, R.M., 2019. Patients' willingness to share digital health and non-health data for research: a cross-sectional study. *BMC medical informatics and decision making*, 19, pp.1-8.
33. Spies, E., Andreu, T., Hartung, M., Park, J. and Kamudoni, P., 2024. Exploring the Perspectives of Patients Living With Lupus: Retrospective Social Listening Study. *JMIR Formative Research*, 8(1), p.e52768.
34. Clusmann, J., Kolbinger, F.R., Muti, H.S., Carrero, Z.I., Eckardt, J.N., Laleh, N.G., Löffler, C.M.L., Schwarzkopf, S.C., Unger, M., Veldhuizen, G.P. and Wagner, S.J., 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1), p.141.
35. Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.B., Aali, A., Bluethgen, C., Paareek, A., Polacin, M., Reis, E.P., Seehofnerova, A. and Rohatgi, N., 2023. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*.
36. Pal, S., Bhattacharya, M., Islam, M.A. and Chakraborty, C., 2023. ChatGPT or LLM in next-generation drug discovery and development: pharmaceutical and biotechnology companies can make use of the artificial intelligence-based device for a faster way of drug discovery and development. *International Journal of Surgery*, 109(12), pp.4382-4384.
37. Zhao, A. and Wu, Y., 2023. Future implications of ChatGPT in pharmaceutical industry: drug discovery and development. *Frontiers in Pharmacology*, 14, p.1194216.
38. Juhi, A., Pipil, N., Santra, S., Mondal, S., Behera, J.K., Mondal, H., Pipil Sr, N., Santra Sr, S. and Behera IV, J.K., 2023. The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus*, 15(3).

39. Ullah, E., Parwani, A., Baig, M.M. and Singh, R., 2024. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology—a recent scoping review. *Diagnostic Pathology*, 19(1), pp.1-9.
40. FDA, U., 2018. Patient-focused drug development: collecting comprehensive and representative input.
41. Tang, L., Sun, Z., Idnay, B., Nestor, J.G., Soroush, A., Elias, P.A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J.F. and Weng, C., 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1), p.158.
1. Ford, E , Shepherd, S , Jones, K , and Hassan, L . Toward an ethical framework for the text mining of social media for health research: a systematic review. *Front Digit Health*. (2021) 2:592237. doi: 10.3389/fdgth.2020.592237
1. Ming Zhang, Jiyu Lu, Jiahao Yang, Jun Zhou, Meilin Wan, Xuejun Zhang, From coarse to fine: Enhancing multi-document summarization with multi-granularity relationship-based extractor, *Information Processing & Management*, Volume 61, Issue 3, 2024
1. 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. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2011, 27730–27744.
1. Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. *arXiv*; 2014. doi: 10.48550/ARXIV.1412.6572
2. Ji Z, Yu T, Xu Y, Lee N, Ishii E, Fung P. Towards Mitigating LLM Hallucination via Self Reflection. In: Bouamor H, Pino J, Bali K, editors. *Findings of the Association for Computational Linguistics: EMNLP 2023 Singapore: Association for Computational Linguistics*; 2023. p. 1827–1843. doi: 10.18653/v1/2023.findings-emnlp.123

Appendix

Table 5: ROUGE Scores and BERTScores obtained for all zero-shot Models without fine-tuning; P values are obtained by comparing to the overall best performing model, i.e., GPT-3.5. Here, B and W indicate the best and worst symptom/treatment number based on each score. The table shows that GPT-3.5 performs best except for ROUGE-2.

Zero-shot Models	ROUGE-1 Mean (SD), B, W	ROUGE-2 Mean (SD), B, W	ROUGE-L Mean (SD), B, W	BERTScore Mean (SD), B, W
Flan-T5	18.91 (± 8.14), 105, 107 ($P < .001$)	3.29 (± 4.55), 105, 4 ($P < .001$)	17.32 (± 7.91), 105, 107 ($P < .001$)	83.34 (± 1.85), 3, 82 ($P < .001$)
GPT-3	32.20 (± 11.04), 45, 17 ($P = .19$)	9.82 (± 8.35), 37, 23 ($P = .63$)	29.10 (± 10.94), 45, 59 ($P = .49$)	87.86 (± 3.00), 45, 17 ($P = .35$)

Zero-shot Models	ROUGE-1 Mean (SD), B, W	ROUGE-2 Mean (SD), B, W	ROUGE-L Mean (SD), B, W	BERTScore Mean (SD), B, W
GPT-3.5	34.00 (± 11.20), 16, 17	10.34 (± 8.96), 16, 28	30.04 (± 11.02), 16, 83	88.21 (± 3.04), 16, 110
Baseline	31.67, ± 17.24 , 6, 67 ($P=.39$)	13.81 , ± 16.5 , 6, 23 ($P=.04$)	29.95, ± 17.10 , 6, 23 ($P=.94$)	85.47 (± 4.06), 104, 17 ($P<.001$)

Table 6: ROUGE-1/2/L and BERTScore for prompt-based GPT-3.5 models with P values obtained by comparing all prompt-based GPT-3.5 models to the best performing GPT-3.5 DSP three-shot model. Here, B and W indicate the best and worst symptom/treatment number based on each score. The table shows that DSP Three-shot performs best although just with a small margin.

Prompt-based GPT-3.5 models	ROUGE-1 Mean (SD), B, W	ROUGE-2 Mean (SD), B, W	ROUGE-L Mean (SD), B, W	BERTScore Mean (SD), B, W
Zero-shot	34.00 (± 11.20), 16, 17 ($P=.55$)	10.34 (± 8.96), 16, 28 ($P=.21$)	30.04 (± 11.02), 16, 83 ($P=.33$)	88.21 (± 3.04), 16, 110 ($P=.98$)
One-shot	32.90 (± 12.24), 7, 59 ($P=.19$)	9.65 (± 9.83), 24, 5 ($P=.08$)	29.01 (± 11.61), 7, 59 ($P=.10$)	88.33 (± 3.10), 16, 59 ($P=.71$)
Three-shot	34.87 (± 11.93), 89, 67 ($P=.98$)	11.62 (± 10.68), 89, 5 ($P=.83$)	31.32 (± 11.80), 89, 67 ($P=.93$)	88.44 (± 2.91), 15, 17 ($P=.50$)
DSP Zero-shot	34.73 (± 11.82), 52, 28 ($P=.92$)	11.39 (± 9.88), 24, 23 ($P=.68$)	31.02 (± 11.35), 16, 83 ($P=.78$)	88.12 (± 2.98), 16, 108 ($P=.82$)
DSP One-shot	34.29 (± 12.32), 6, 17 ($P=.68$)	10.90 (± 10.69), 37, 28 ($P=.45$)	30.64 (± 12.19), 105, 17 ($P=.57$)	88.17 (± 3.04), 104, 108 ($P=.94$)
DSP Three-shot	34.91 (± 11.42), 89, 17	11.91 (± 10.68), 89, 28	31.50 (± 11.63), 89, 59	88.20 (± 2.79), 15, 119

CoT Zero-shot	31.45 (± 10.84), 16, 59 ($P=.01$)	9.23 (± 7.72), 16, 28 ($P=.02$)	28.07 (± 10.55), 113, 59 ($P=.01$)	87.28 (± 2.64), 16, 17 ($P=.009$)
------------------	---	---	--	---

Table 7: ROUGE-1/2/L and BERTScore for Flan-T5 Model variants with P values obtained by comparing all Flan-T5 models to the fine-tuned Flan-T5 model. Here, B and W indicate the best and worst symptom/treatment number based on each score. The table shows that the fine-tuned model performs better than all non-fine-tuned models no matter what prompting strategy is used.

Flan-T5 Model Variants	ROUGE-1 Mean (SD), B, W	ROUGE-2 Mean (SD), B, W	ROUGE-L Mean (SD), B, W	BERTScore Mean (SD), B, W
Zero-shot Flan-T5	19.59 (± 5.30), 114, 121 ($P<.001$)	2.82 (± 1.52), 115, 121 ($P<.001$)	18.40 (± 5.31), 114, 121 ($P<.001$)	83.69 (± 1.57), 125, 122 ($P<.001$)
One-shot Flan-T5	21.08 (± 7.68), 114, 124 ($P<.001$)	2.88 (± 2.63), 116, 117 ($P=.007$)	18.66 (± 7.53), 116, 124 ($P<.001$)	83.87 (± 2.00), 118, 120 ($P<.001$)
Three-shot Flan-T5	19.96 (± 7.82), 117, 124 ($P<.001$)	3.23 (± 2.97), 116, 114 ($P=0.01$)	17.81 (± 6.96), 117, 124 ($P<.001$)	83.27 (± 1.76), 117, 121 ($P<.001$)
CoT Zero-shot Flan-T5	20.00 (± 9.26), 127, 119 ($P<.001$)	2.75 (± 2.49), 127, 114 ($P<.001$)	18.31 (± 8.38), 127, 119 ($P<.001$)	83.36 (± 2.52), 118, 119 ($P<.001$)
DSP Zero-shot Flan-T5	24.84 (± 10.29), 117, 119 ($P=.002$)	6.49 (± 7.28), 117, 114 ($P=.16$)	22.81 (± 9.77), 117, 119 ($P=.003$)	84.05 (± 2.98), 117, 119 ($P<.001$)
DSP One-shot Flan-T5	18.96 (± 6.89), 127, 116 ($P<.001$)	3.32 (± 2.85), 117, 116 ($P=.09$)	17.83 (± 6.90), 127, 116 ($P<.001$)	83.30 (± 2.10), 117, 116 ($P<.001$)
DSP Three-shot Flan-T5	23.60 (± 6.25), 116, 124 ($P=.009$)	4.28 (± 4.22), 117, 124 ($P=.44$)	21.08 (± 5.77), 116, 124 ($P=.02$)	84.38 (± 1.86), 118, 125 ($P<.001$)
Fine-tuned Flan-T5	30.17 (± 6.42), 117, 119	7.69 (± 4.16), 117, 124	27.53 (± 5.23), 117, 119	86.52 (± 1.12), 115, 124

Tables 8 and 9 show the average ROUGE-1/2/L and BERTScore for 124 summaries generated by using different prompting techniques on Flan-T5 and GPT-3. In both models, DSP Three-shot prompting shows better performance when compared to other prompting techniques.

Table 8: ROUGE-1/2/L and BERTScore obtained for Prompt-based Flan-T5 models in the experiment with P values obtained by comparing all prompt-based Flan-T5 models to the best performing Flan-T5 DSP Three-shot model. Here B and W indicate the best and worst symptom/treatment number based on each score. The table shows that DSP Three-shot performs best.

Prompt-based Flan-T5 models	ROUGE-1 Mean (SD),B,W	ROUGE-2 Mean (SD), B,W	ROUGE-L Mean (SD),B, W	BERTScore Mean (SD),B, W
Zero-shot	18.91 (± 8.14), 105, 107 ($P < .001$)	3.29 (± 4.55), 105, 4 ($P < .001$)	17.32 (± 7.91), 105, 107 ($P < .001$)	83.34 (± 1.85), 3, 82 ($P = .63$)
One-shot	19.83 (± 8.76), 3, 45 ($P = .005$)	3.84 (± 5.06), 3, 21 ($P = .008$)	18.29 (± 8.26), 3, 45 ($P = .005$)	83.28 (± 2.07), 37, 112 ($P = .52$)
Three-shot	20.50 (± 8.86), 37, 59 ($P = .02$)	3.95 (± 5.33), 37, 11 ($P = .01$)	18.44 (± 8.07), 37, 59 ($P = .007$)	83.40 (± 1.96), 29, 76 ($P = .80$)
DSP Zero-shot	22.54 (± 8.77), 117, 67 ($P = .60$)	5.35 (± 5.99), 117, 21 ($P = .42$)	20.57 (± 8.40), 117, 67 ($P = .41$)	83.54 (± 2.27), 117, 124 ($P = .79$)
DSP One-shot	21.52 (± 8.37), 96, 116 ($P = .16$)	5.06 (± 5.56), 96, 10 ($P = .25$)	19.85 (± 8.21), 96, 116 ($P = .14$)	83.42 (± 2.17), 3, 24 ($P = .89$)
DSP Three-shot	23.14 (± 9.69), 29, 59	6.05 (± 7.73), 29, 20	21.50 (± 9.57), 29, 59	83.46 (± 2.28), 29, 52
CoT Zero-shot	18.34 (± 8.84), 105, 119 ($P < .001$)	2.95 (± 4.16), 105, 5 ($P < .001$)	16.76 (± 8.32), 105, 119 ($P < .001$)	82.90 (± 2.08), 118, 20 ($P = .04$)

Table 9: ROUGE-1/2/L and BERTScore obtained for Prompt-based GPT-3 models in the experiment with P values obtained by comparing all prompt-based GPT-3 models to the best performing GPT-3 DSP Three-shot model. Here B and W indicate the best and worst symptom/treatment number based on each score. The table shows that DSP Three-shot performs best.

Prompt-based GPT-3 models	ROUGE-1 Mean (SD),B, W	ROUGE-2 Mean (SD),B, W	ROUGE-L Mean (SD),B, W	BERTScore Mean (SD),B, W
Zero-shot	32.20 (± 11.04),	9.82 (± 8.35),	29.10	87.86 (± 3.00),

	45, 17 ($P=.14$)	37, 23 ($P=.61$)	(± 10.94), 45, 59 ($P=.40$)	45, 17 ($P=.67$)
One-shot	33.00 (± 11.11), 89, 83 ($P=.38$)	9.50 (± 7.88), 24, 28 ($P=.39$)	28.93 (± 10.25), 89, 83 ($P=.30$)	88.31 (± 2.90), 16, 108 ($P=.39$)
Three-shot	33.91 (± 9.37), 6, 17 ($P=.83$)	9.19 (± 6.51), 37, 23 ($P=.20$)	29.47 (± 8.65), 68, 59 ($P=.51$)	88.06 (± 2.71), 16, 18 ($P=.86$)
DSP Zero-shot	32.43(± 10.95), 52, 67 ($P=.21$)	9.52 (± 8.47), 104, 4 ($P=.44$)	28.97 (± 11.00), 52, 67 ($P=.37$)	87.54 (± 2.79), 16, 23 ($P=.19$)
DSP One-shot	33.03 (± 11.34), 37, 67 ($P=.42$)	9.68 (± 8.67), 37, 23 ($P=.53$)	29.65 (± 11.12), 37, 83 ($P=.68$)	88.00 (± 2.91), 16, 19 ($P=.99$)
DSP Three-shot	34.20 (± 10.51), 37, 17	10.34 (± 8.09), 37, 5	30.23 (± 10.01) , 37, 17	88.00 (± 2.64), 16, 17
CoT Zero-shot	31.74 (± 11.72), 105, 67 ($P=.08$)	10.27 (± 9.89), 105, 5 ($P=.94$)	28.78 (± 11.55), 105, 67 ($P=.30$)	87.45 (± 2.98), 15, 60 ($P=.12$)

The prompt templates for each prompting technique used in this research are given below. In the given prompt templates $\{\text{text}\}$ denotes the patient comment regarding a symptom or treatment drawn from the example set, $\{\text{hint}\}$ denotes hints which are a combination of keywords that are extracted by KeyBERT with KeyphraseVectorizers and frequent terms that occur in each patient comment, $\{A\}$ denotes the generative language model generated answer for each question and $\{\text{summary}\}$ denotes the reference summary of the corresponding patient comment.

Zero-shot prompt template:

Write a short summary for this patient comment: {text}
Summary:

One-shot prompt template:

Given a patient comment, write a short summary for the patient comment.

Write a short summary for this patient comment: {text}
Summary: {summary}

Write a short summary for this patient comment: {text}
Summary:

Figure 4: One-shot prompt template used in the experiments.

Three-shot prompt template:

Given a patient comment, write a short summary for the patient comment.

Write a short summary for this patient comment: {text}
Summary: {summary}

Write a short summary for this patient comment: {text}
Summary: {summary}

Write a short summary for this patient comment: {text}
Summary: {summary}

Write a short summary for this patient comment: {text}
Summary:

Directional Stimulus prompt - Zero-shot template:

Write a short summary based on the provided hint for this patient comment: {text}
Hint: {hint}
Summary:

Directional Stimulus prompt - One-shot template:

Given a patient comment and hint, write a short summary for the patient comment based on the hint.

Write a short summary based on the provided hint for this patient comment: {text}

Hint:{hint}

Summary: {summary}

Write a short summary based on the provided hint for this patient comment: {text}

Hint: {hint}

Summary:

Directional Stimulus prompt - Three-shot template:

Given a patient comment and hint, write a short summary for the patient comment based on the hint.

Write a short summary based on the provided hint for this patient comment: {text}

Hint:{hint}

Summary: {summary}

Write a short summary based on the provided hint for this patient comment: {text}

Hint:{hint}

Summary: {summary}

Write a short summary based on the provided hint for this patient comment: {text}

Hint:{hint}

Summary: {summary}

Write a short summary based on the provided hint for this patient comment: {text}
Hint: {hint}
Summary:

Chain-of-Thought Zero-shot prompt template - Flan-T5:

What is the main health or medical related topic discussed in this patient comment?: {A}
What are the important informations mentioned in this patient comment?: {A}
What are the symptoms and diseases mentioned in this patient comment?: {A}
What are the recovery methods or treatments mentioned in this patient comment?: {A}
Let's integrate the above information and summarize this patient comment: {text}
Summary:

Chain-of-Thought Zero-shot prompt template - GPT-3 & GPT-3.5:

1. The main health or medical related topic {A}.
2. The important information mentioned {A}.
3. The diseases mentioned {A}.
4. The recovery methods or treatments mentioned {A}.

Let's integrate the above information and summarize the patient comment: {text}
Summary: