Liu *et al.*
BMC Medical Informatics and Decision Making (2024) 24:75
https://doi.org/10.1186/s12911-024-02481-8

BMC Medical Informatics and Decision Making

RESEARCH **Open Access**

Exploring the potential of ChatGPT in medical dialogue summarization: a study on consistency with human preferences

Yong Liu[1], Shenggen Ju[1*†] and Junfeng Wang[1†]

Abstract

Background Telemedicine has experienced rapid growth in recent years, aiming to enhance medical efficiency and reduce the workload of healthcare professionals. During the COVID-19 pandemic in 2019, it became especially crucial, enabling remote screenings and access to healthcare services while maintaining social distancing. Online consultation platforms have emerged, but the demand has strained the availability of medical professionals, directly leading to research and development in automated medical consultation. Specifically, there is a need for efficient and accurate medical dialogue summarization algorithms to condense lengthy conversations into shorter versions focused on relevant medical facts. The success of large language models like generative pre-trained transformer (GPT)-3 has recently prompted a paradigm shift in natural language processing (NLP) research. In this paper, we will explore its impact on medical dialogue summarization.

Methods We present the performance and evaluation results of two approaches on a medical dialogue dataset. The first approach is based on fine-tuned pre-trained language models, such as bert-based summarization (BERTSUM) and bidirectional auto-regressive Transformers (BART). The second approach utilizes a large language models (LLMs) GPT-3.5 with inter-context learning (ICL). Evaluation is conducted using automated metrics such as ROUGE and BERTScore.

Results In comparison to the BART and ChatGPT models, the summaries generated by the BERTSUM model not only exhibit significantly lower ROUGE and BERTScore values but also fail to pass the testing for any of the metrics in manual evaluation. On the other hand, the BART model achieved the highest ROUGE and BERTScore values among all evaluated models, surpassing ChatGPT. Its ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore values were 14.94%, 53.48%, 32.84%, and 6.73% higher respectively than ChatGPT's best results. However, in the manual evaluation by medical experts, the summaries generated by the BART model exhibit satisfactory performance only in the "Readability" metric, with less than 30% passing the manual evaluation in other metrics. When compared to the BERTSUM and BART models, the ChatGPT model was evidently more favored by human medical experts.

Conclusion On one hand, the GPT-3.5 model can manipulate the style and outcomes of medical dialogue summaries through various prompts. The generated content is not only better received than results from certain human experts but also more comprehensible, making it a promising avenue for automated medical dialogue summarization. On the other hand, automated evaluation mechanisms like ROUGE and BERTScore fall short in fully assessing

†Shenggen Ju and Junfeng Wang contributed equally to this work.

*Correspondence:
Shenggen Ju
jsg@scu.edu.cn
Full list of author information is available at the end of the article

the outputs of large language models like GPT-3.5. Therefore, it is necessary to research more appropriate evaluation criteria.

Keywords Internet Healthcare, Large language models, ChatGPT, Automated medical consultation, Medical dialogue summarization

Background

As healthcare evolves towards a patient-centered delivery model, online searching and accessing health information can fulfill patients' needs for prognosis and treatment information [1, 2]. Furthermore, with the rapid development of "Internet plus Healthcare" online consultation platforms are emerging, allowing doctors to diagnose diseases and provide relevant medical advice through remote conversations with patients. On the one hand, this enhances the efficiency of the healthcare system and alleviates some of the burden on medical professionals, enabling them to invest more energy in improving patient care and minimizing time spent on irrelevant matters [3, 4]. On the other hand, it enables effective patient screening, maintains social distancing, and protects clinical doctors and communities from infections, while still offering personalized healthcare and medical services [5]. During the COVID-19 pandemic, influenced by policies and the pandemic's impact, the demand for online consultations has rapidly increased. This has irrevocably altered the status of telemedicine in the U.S. healthcare system and has been widely adopted across global healthcare systems [6]. Summarizing conversations on remote medical platforms can bring about several benefits. For instance, Both doctors and patients can refer to important parts or conclusions from past interactions. This not only allows patients to quickly access the results they are concerned about but also enables doctors to learn from the experiences and approaches of other medical professionals when dealing with similar issues. Medical text summarization algorithms related to medical dialogue summaries are techniques that automatically extract key information from various medical data sources such as medical literature, electronic health records, and medical dialogue, and generate concise summaries. These algorithms mainly include two types: extractive summarization and abstractive summarization.

Extractive summarization selects content based on importance or keywords in the text. The generation process does not involve creating new sentences or phrases; it simply selects and combines existing content from the original text to generate a summary. It often treats summarization as a sequence labeling task, where each sentence is labeled with a binary classification tag of "yes" or "no", and the summarization process can be viewed as the selection of sentence classification labels. BERTSUM is a text summarization model based on bidirectional encoder representation from Transformers (BERT). It aims to leverage the powerful representation capabilities of BERT to generate concise summaries from input text [7]. A hierarchical encoder-tagger model enhanced with a memory module was proposed to identify important utterances in dialogues between patients and doctors, thereby accomplishing the task of medical dialogue summarization [8]. However, this method selects essential sentences from the original text to form concise summaries, maintaining interpretability and accuracy, however, it lacks the ability to generate new sentences, potentially resulting in less smooth and coherent summaries [8].

The abstractive summary uses natural language generation techniques to create new sentences or phrases by understanding the semantics and context of the original text in order to generate a summary. Abstractive summarization can typically express information more freely, rather than relying solely on the extraction of content from the original text. Kundan et al. [9] proposed a bidirectional LSTM-based encoder-decoder model with attention. It is used to extract important phrases related to each section of the summary and concatenate these relevant phrases together to generate a summary sentence for each cluster. The BART model combines the encoder-decoder architecture of Transformer with a pre-training task involving denoising autoencoders. This structure helps the model capture complex relationships between input data in text generation tasks [10]. A simple yet general two-stage fine-tuning method is proposed to deal with input length limitations of the model, enabling the step-by-step generation of medical dialogue summaries [11]. George et al. [12] proposed a sequence-to-sequence architecture for summarizing medical dialogues by integrating medical domain knowledge from the Unified Medical Language System (UMLS). Medical concepts from the references are encoded to distinguish important medical concepts, and combining an end-to-end approach to explicitly model a switching variable, induce a mixed model of copying, generating, and negation to obtain medical dialogue summaries [13]. The abstractive summary algorithm uses deep learning techniques such as RNN or Transformer to generate new, fluent, and coherent summaries, and ensures accuracy while avoiding the generation of unreasonable summaries when dealing with complex medical texts [11, 14]. The success of Transformer is attributed to their high degree of parallelism and self-attention mechanism. Building upon this foundation, the BERT model improved the architecture and achieved universal language representations through unsupervised pre-training on large-scale corpora [15]. This study inspired a great deal of subsequent work, establishing the "pre-training and fine-tuning" learning paradigm, and introducing different architectures such as BART [16] and GPT-2 [17] that can be used to fine-tune downstream tasks. As the scale of parameters continues to increase to hundreds of billions, and training on massive data, large language models such as GPT are eventually generated, including GPT-3 and GPT-4 [18].

In recent years, LLMs provides opportunities for instructional fine-tuning through reinforcement learning from human feedback (RLHF) [19] to adapt to a variety of NLP tasks while aligning the model with human intent. They excel in areas such as education, healthcare, text generation and human-computer interaction. Summarization based on LLMs employs ICL and pre-trained language models (PLMs) to generate clinical notes, the results show that the ICL-based approach is as well-received as human-written notes. This makes it a promising approach for automatically generating notes in medical dialogues [20]. Expert validation demonstrates that clinical notes generated by ICL in GPT-4 outperform all traditional fine-tuned models [21]. The applicability of large language models in radiology report summarization tasks was explored by optimizing input prompts based on a small number of existing samples and an iterative approach [22]. ChatGPT allows doctors to input specific information and medical concepts related to patients and generate formal patient discharge summaries. Automating this process can reduce the workload of junior doctors, giving them more time to provide patient care and seek training opportunities [23].

In this paper, We utilized two different approaches to generate medical dialogue summaries on a publicly available medical conversation dataset. Method A involves fine-tuning PLMs such as BERTSUM [7] and BART [16], while Method B utilizes the large language model ChatGPT based on ICL [24], whose processing flow is shown in Fig. 1. For Method B, we first fine-tune the relevant parameters of the ChatGPT model and then explore using the model's prompt engineering functionality to generate medical dialogue summaries, and finally measure the ability of ChatGPT to generate medical dialogue summaries by automatic and human evaluation criteria.

Methods

Model

Transformer

The "Transformer" model has become a foundational deep learning architecture for natural language processing [25]. As shown in Fig. 2, the Transformer model is composed of encoder and decoder layers. Each of the encoder layer has a multi-head self-focusing and position-wise feed-forward network (FFN) sub-layer. The

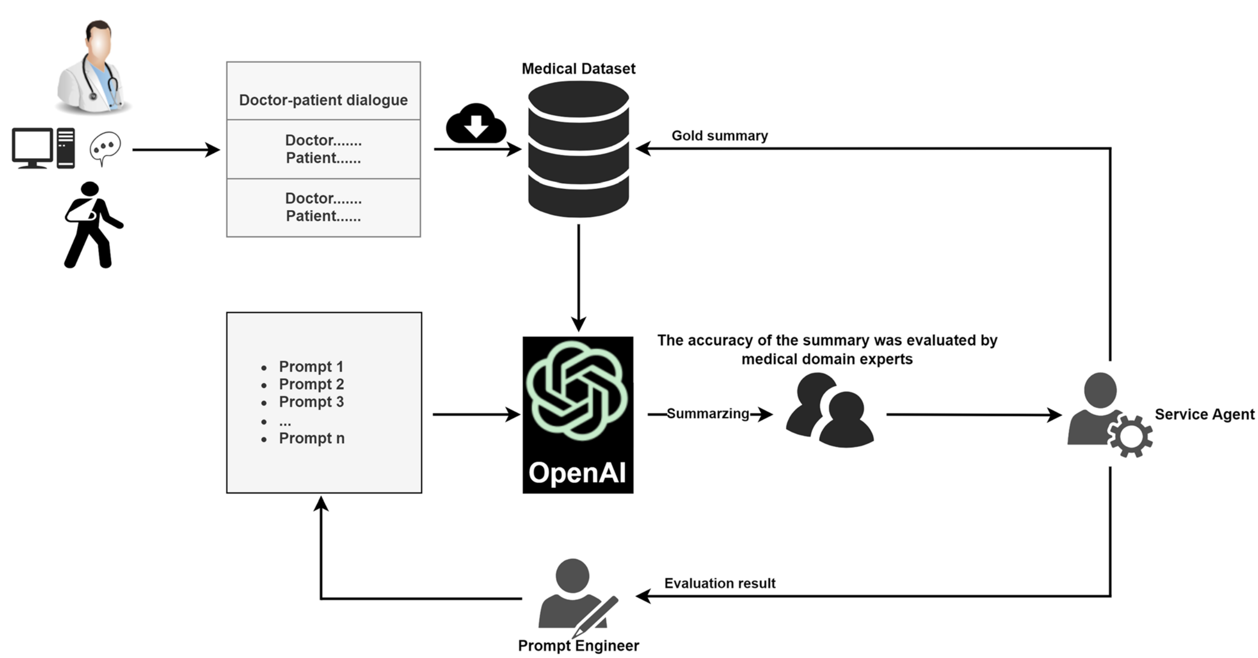


Fig. 1 The processing flow of generating summaries from doctor-patient dialogues using ChatGPT

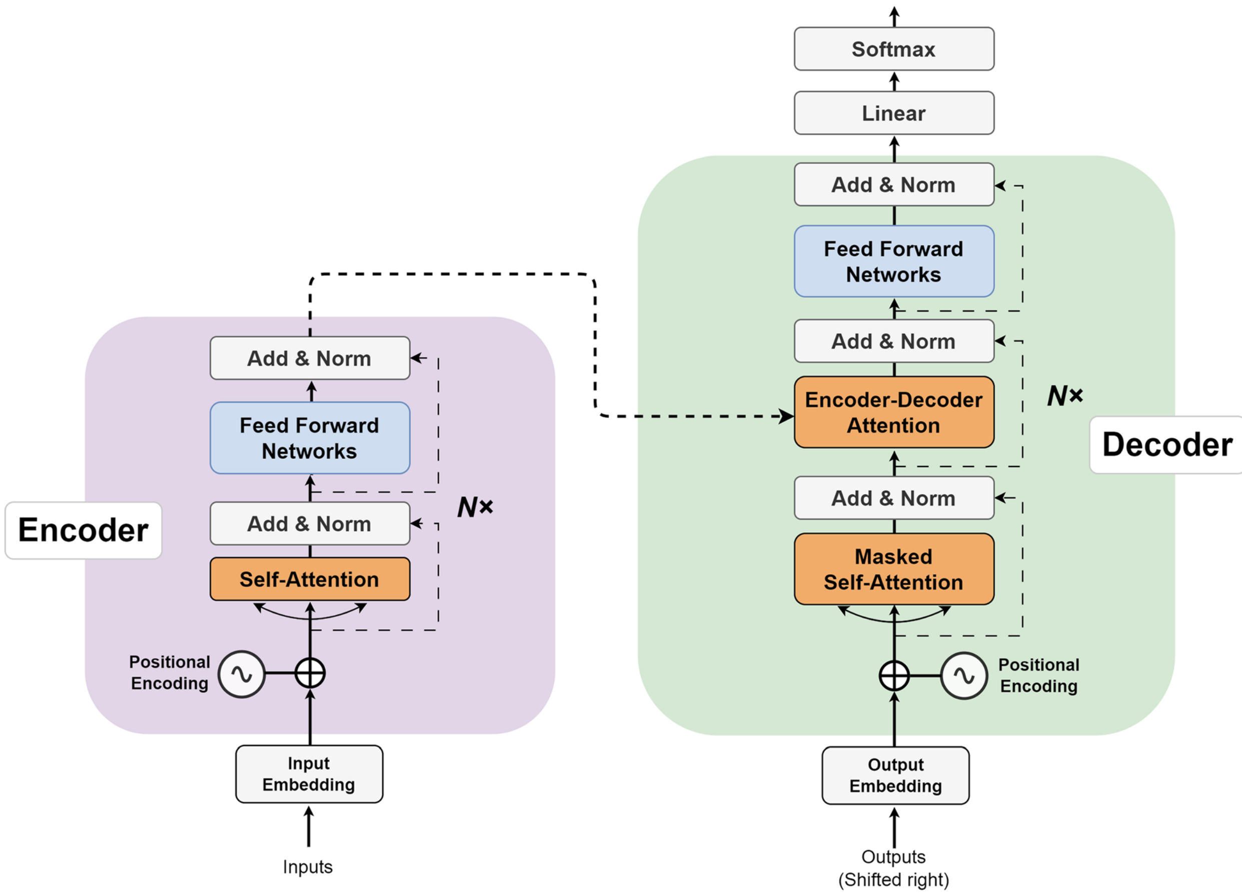


Fig. 2 The Transformer - model architecture. The creation of this figure is based on Fig. 1 in the paper [25]

multi-head attention mechanism utilizes parallel scaled dot-product attention functions to focus on different subspaces and positions in the input data. By calculating the attention between query (Q), keys (K), and values (V), the results are as follows:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{d_k}\right)V \tag{1}$$

In Eq. (1), a set of queries , keys and values are packed together into a matrices Q, K and V. In addition, d_k usually refers to the dimension of the key vector, which can be used as a scaling factor to avoid excessive dot products. ReLU activation is used in the FFN sub-layer. In addition, layer normalization and a residual connection link the two sub-layers and can be used to tackle gradient issues , thus, stable network training can be obtained. Each decoder layer includes three sub-layers: an FFN sub-layer and two attention sub-layers. The decoder self-attention sub-layer uses a mask function to prevent attending to unseen future tokens. The encoder-decoder attention layer enables the decoder to focus on essential parts of the source sequence and capture the encoder-decoder relationship.

Given an input sequence of symbol representations $X = [x_1, x_2, ..., x_n]$ and a real output sequence $Y = [y_1, y_2, ..., y_m]$. We assume that the last position of each input sequence is a special "[END]" flag. The encoder maps X to a sequence of continuous representations $H = [h_1, h_2, ..., h_n]$. Given H, the decoder then generates a predicted output sequence $\hat{Y} = [\hat{y_1}, \hat{y_2}, ..., \hat{y_k}]$. Therefore, during training, the model minimizes the cross-entropy loss between the predicted sequence $\hat{Y}$ and the real output sequence Y, as shown in Eq. (2), where k represents the number of target sequences. At each step, the model is auto-regressive [26], which means that it utilizes the previously generated symbols as additional input when generating the next symbol in the sequence. This enables the model to contextually understand and produce coherent outputs in a step-by-step manner.

$$L\left(\hat{Y}, Y\right) = -\sum_{i=1}^{m} \hat{y}_i log y_i \tag{2}$$

The Transformer's impact on natural language processing has been profound, inspiring the development of modern NLP models such as BERT [15], GPT [17], RoBERTa [27], and T5 [28], all built on the Transformer architecture.

BERT

BERT is an important language model based on the Transformer architecture, it is trained to learn general language representations and capture contextual semantics, which has had a profound impact on NLP research and applications [29–31]. As shown in Fig. 3, the basic BERT structure is made up of multiple layers of Transformer and includes two pre-training tasks: mask language model (MLM) and next sentence prediction (NSP). Taking NSP as an example, the main elements are divided three parts as follows:

- Input layer. For a given input text that has undergone masking and is represented as $x = [x_1 x_2 \ldots x_n]$ and $x' = [x'_1 x'_2 \ldots x'_m]$, the following processing results in the BERT input representation e.

$$X = [CLS]x_1x_2 \ldots x_n[SEP]x'_1x'_2 \ldots x'_m[SEP] \tag{3}$$

$$e = InputRepresentation\ (X) \tag{4}$$

 In Eq. (3), [CLS] represents the special token marking the beginning of a text sequence, and [SEP] represents a separator marker between text sequences.

- BERT encoder layer. In this layer, the input representation e is encoded by L layers Transformer to obtain a contextual semantic representation of the input text:

$$h = Transformer\ (e) \tag{5}$$

 where $h \in \mathbb{R}^{N \times d}$, N represents the maximum length of the sequence, and d represents the hidden layer dimension of BERT.

- The output layer. In the NSP task, BERT uses the hidden layer of the [CLS] position as the semantic representation of the context consisting of the first component h_0 of h, and finally predicts the classification probability P of the input text through a fully connected layer:

$$P = Softmax\left(h_0 W^p + b^0\right) \tag{6}$$

 where $P \in \mathbb{R}^2$, $W^p \in \mathbb{R}^{d \times 2}$ represents the weight of the fully connected layer, and $b^0 \in \mathbb{R}^2$ represents the bias of the fully connected layer. Classification probability P is used to calculate cross-entropy loss with the real label y. The final model parameters are updated based on this loss.

The application of BERT in medical document summarization accelerates the acquisition [32, 33], processing and application of medical information [34, 35], improving the efficiency and accuracy of healthcare and medical research [36]. It brings a broader development space for

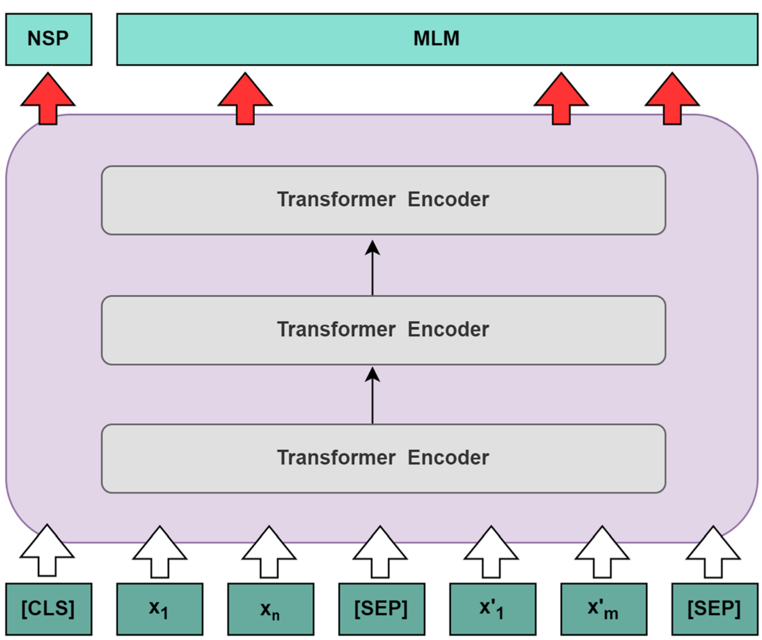


Fig. 3 The overview architecture of the BERT model. The creation of this figure is based on Fig. 1 in the paper [15]

the medical field, and is expected to promote the innovation of medical information processing and application in the future [37].

BERTSUM

BERTSUM is a text summarization model based on BERT which aims to generate concise summaries from input text by leveraging BERT's powerful representation capabilities. The model utilizes BERT's encoder to obtain semantic vector representations for sentence-level segments in the input text. The training of BERTSUM occurs in two stages: pre-training of the BERT model on unsupervised tasks and supervised fine-tuning with datasets containing human summaries. During fine-tuning, the model optimizes the similarity between generated and original summaries as the loss function [7].

As shown in Fig. 4, in the context of encoding multiple sentences, an [CLS] token is inserted before each sentence, and a [SEP] token is inserted at the end of each sentence. In vanilla BERT, [CLS] is used to aggregate features from a single sentence or a pair of sentences. Therefore, by using multiple [CLS] tokens to fine-tune the model and based on these tokens, it is possible to obtain sentence features in ascending order. The BERT sentence vectors undergo additional summarization-specific layers to capture document-level features for summary extraction. The resulting values are then passed through the sigmoid function, which maps them to a range between 0 and 1. Therefore, each sentence is assigned a predicted score $\hat{Y}$.

$$\hat{Y} = \sigma(W_o T_i + b_o) \tag{7}$$

The model's loss is the binary classification entropy between $\hat{Y}$ and the gold label Y.

BART

BART is an NLP pre-trained model proposed by Facebook AI for pre-training the bidirectional and auto-regressive combined model Transformers [16]. The main idea is to combine the encoder-decoder structure of Transformer with the pre-training task of denoising auto-encoder. BART uses an encoder-decoder structure similar to Transformer. The encoder is responsible for mapping the input sequence to an intermediate representation, and the decoder then maps this intermediate representation back to the original input space. This structure can help the model capture complex relationships between inputs. In the text generation task, the input of the encoder is the input text as the condition, and the decoder generates the corresponding target text in an auto-regressive way, as shown in Fig. 5.

The BART model has acquired a substantial amount of basic language knowledge during the pre-training phase, so during downstream tasks (such as text classification, named entity recognition, question answering system, document summarization, etc.), we only need to fine-tune the model, without needing to train

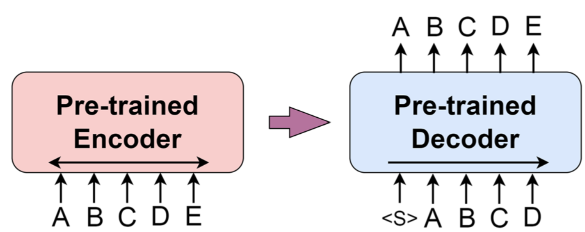


Fig. 5 Example of a task used by the BART model for text generation. The creation of this figure is based on Fig. 3 in the paper [16]

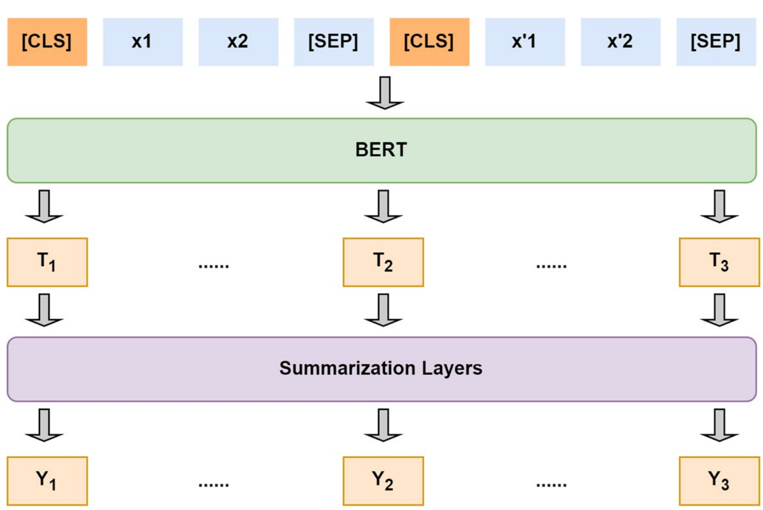


Fig. 4 The overview architecture of the BERTSUM model. The creation of this figure is based on Fig. 1 in the paper [7]

the model from scratch. This greatly saves training time and improves the performance of the model.

ChatGPT and prompt

ChatGPT

ChatGPT is a powerful language model developed by OpenAI based on the Transformer architecture, which can be implemented in three steps [19]. Step 1, Collect demonstration data and train a supervised policy. Step 2, Collect comparison data and train a reward model. Step 3, Optimize the reward model using proximal policy optimization (PPO) [38]. Figure 6 illustrates the relevant process of Step 2, where annotators annotate the data in the candidate dataset according to their respective standards and manually rank them based on the scores. Then, they input the rankings into a reward model to predict the preferences of the manual annotations.

The loss function used during training is shown in formula (8).

$$Loss(\theta) = -\frac{1}{\binom{K}{2}} E\left(x, y_w, y_l\right) \sim D\left[log\left(\sigma\left(r_\theta\left(x, y_w\right) - r_\theta\left(x, y_l\right)\right)\right)\right] \quad (8)$$

where $r_\theta\left(x, y\right)$ is the scalar output of the reward model for prompt x and completion y with parameters θ , y_w is the preferred completion out of the pair of y_w and y_l, and D is the comparison dataset.

As a generative model, ChatGPT can produce text sequences given an initial prompt. It comes in different versions, with each new iteration being more capable and better at handling complex language tasks [39]. ChatGPT finds applications in various fields [40], including tutoring and education [41], translation [42], healthcare [43], and medicine [44–46].

Regarding medical text summarization, users can utilize ChatGPT to understand and condense lengthy medical reports or research papers [47]. By using a relevant portion of the text as a prompt, the model can provide a concise summary tailored to the user's needs [48]. For example, a user studying heart disease could prompt the model to generate a simplified summary of coronary artery disease and its causes.

Prompts for large language models

In ChatGPT or other GPT models, "prompt" refers to the initial text or request entered into the model. This can be a question, part of a sentence, or even a word. The model generates or continues text based on this initial prompt, with the goal of producing fast responses that are consistent in terms of grammar, context, and style. It achieves this by relying on the language patterns and associations learned during the pre-training phase, as well as the more specific guidance acquired during the fine-tuning phase.

By optimizing the design of prompt, users can better guide models to produce outputs that meet their specific needs. We try to provide some reference content as follows:

- Define the task or objective. Clearly specify the task or objective you want the model to accomplish. This could be answering questions, generating articles,etc.
- Understand the model's capabilities. Familiarize yourself with the limitations and strengths of the model. Different models perform differently on tasks, and understanding their capabilities helps guide them effectively.
- Choose appropriate length and format. Determine the length and format of the prompt. Sometimes, concise prompts are more effective, but for certain tasks, a more detailed description or context may be necessary.
- Grammar and format. Ensure the prompt is grammatically correct and adheres to the input format expected by the model. This increases the likelihood of the model understanding and generating correct output.

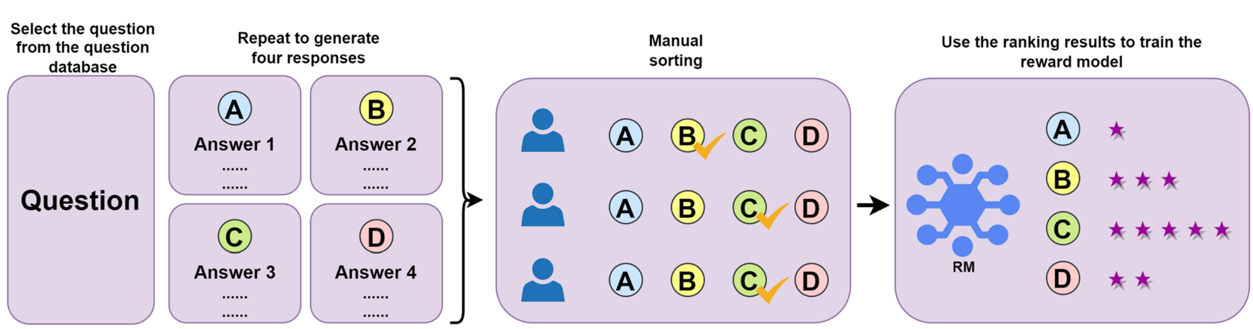


Fig. 6 Training Process of the reward model (RM)

- Provide relevant information. If specific background or contextual information is required, make sure to include it in the prompt. This helps the model better understand the task or question.
- Iterate for optimization. Iterate through different prompts, observe the model's responses, and adjust based on the results. This is an iterative process that gradually improves the model's performance.
- Evaluate and adjust. Assess whether the generated text aligns with expectations and adjust or improve prompts as needed. Continuous evaluation of outputs guides the optimization process.

Follow the above description, users can better guide models to produce outputs that meet their specific needs. Overall, prompt is an important factor in driving GPT models to produce specific outputs, and users can carefully design prompt to get the best model output [49].

Experiments

Dataset

We utilized the data provided by the School of Data Science at Fudan University, which was constructed under the guidance of medical experts from Fudan University Medical School and named IMCS-V2. This dataset has collected authentic online medical dialogues and subjected them to multi-level human annotations. The aim is to facilitate open evaluation against the Chinese biomedical language understanding evaluation (CBLUE) benchmark and thereby advance the fields of intelligent healthcare and medical language understanding. The IMCS-V2 dataset comprises 4,116 medical-patient dialogue samples that have undergone meticulous annotations, and detailed statistical data are presented in the following Table 1. In addition, this dataset encompasses 10 pediatric diseases, and the disease distribution is shown in Fig. 7.

The dataset used in this paper is a publicly available dataset designed for a medical natural language processing competition. The task involves generating medical reports from multi-turn doctor-patient dialogues. However, due to the competition's requirements, the summary portions of the test set have been omitted. This limitation prevents us from comparing it with the summaries generated by ChatGPT. Therefore, only (4116 - 811)=3305 samples out of the original 4116 samples are available for the final training and testing. We migrate the first 500 samples (sorted by file name) from the valid set as the new test set, and then combine the first 167 samples from the training set (sorted by file name) with the remaining samples from the valid dataset to form a new valid dataset consisting of 500 samples. In the end, train set has 2305 samples, valid and test sets have 500 samples respectively. Since the samples distribution in the original train and valid sets were randomly generated by the organizer, we did not carry out random selection in the process of data

Table 1 Detailed statistics on the IMCS-V2 medical dialogue dataset

Statistical indicators	Value
Total Diseases	10
Total Dialogues	4116
Total Sentences	164731
Average Sentences/Per Dialog	40
Average Words/Per Dialog	523

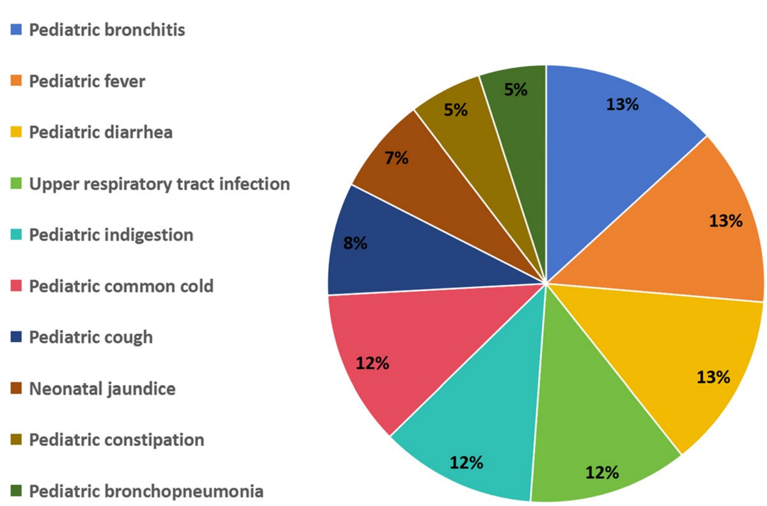


Fig. 7 The proportion distribution of the 10 pediatric diseases in the IMCS-V2 medical dialogue dataset

migration, but only extracted according to the order of file names.

Experimental environment and model parameter settings

In this paper, all the experiments were conducted on two NVIDIA GeForce RTX 3090 GPUs with 24GB of memory and using the python language on the PyCharm platform. The relevant environment settings required for the experiment are shown in Table 2. The data is divided into a training set of 2305, a verification set of 500 and a test set of 500. In this paper, abstracts are divided into two types: extractive summarization and abstractive summarization. BERTSUM model is the representative of extractive summarization, while BART model and ChatGPT are the representatives of abstractive summarization. The training parameters of BERTSUM model are shown in Table 3, and the training parameters of BART model are shown in Table 4.

Prompt settings for summarization

Figures 8, 9 and 10 show prompt settings related to ChatGPT model. The prompts in this paper are divided into two types: simple type, denoted as Prompt_S, as shown in Fig. 8, and technical type, denoted as Prompt_T, as shown in Fig. 9. In Fig. 8, Prompt_S does not use complicated prompts, but indicates "Medical examination names" in the "Auxiliary examination" part, and "The most relevant diagnostic name" in the "Diagnosis" part, in the hope that ChatGPT can return relevant results.

Since this article is based on research using a Chinese medical dialogue dataset, the prompts used by ChatGPT are written in Chinese. Actually, ChatGPT itself supports multiple languages, so you can also write prompts in English, and simply include an additional instruction for ChatGPT to provide medical dialogue summaries in Chinese. Please note that there will be several instances of Chinese content in this text, along with corresponding English explanations, just to make it clearer for readers who use different languages. In Fig. 9, taking advantage of the idea of task decomposition, we use Prompt_T to split the summary into six parts, each with related more detailed sub-prompts. For example, the "Chief complaint: Not exceeding 20 words, including symptoms such as {*ls_symptom*} as an example, generated based on the doctor-patient dialogue, and the duration of symptoms.". Symptoms are stored in the variable {*ls_symptom*}, which mainly refers to symptoms of childhood diseases related to the dataset, including "fever", "cough", "sore throat", "runny nose", "abdominal pain", "diarrhea", etc. Please refer to Fig. 10 for detailed explanations. The purpose of this approach is to try to give ChatGPT, in this way, a contextual example of similar symptoms of the disease, so that it can give a better description of the patient's symptoms. Similarly, in the "Auxiliary examination" part, "generate the most relevant medical examination names, using {*ls_aux_test*} as an example". The variable {*ls_aux_test*} stores some medical examination names, such as "complete blood count", "urinalysis", "stool routine", "chlamydia", "liver function", "kidney function", etc. Eventually, ChatGPT is expected to give a better name for the auxiliary examinations performed by the patient based on the content of this variable {*ls_aux_test*}. For the rest of Fig. 9, some give specific examples, such as the "Diagnosis" part, and some give relevant explanations, such as the "Present medical history" part. For details, please refer to

Table 3 Hyper-parameters of BERTSUM, in the case of multiple candidate parameter values, the ultimately chosen parameter value is displayed in bold

Parameters	Values
encoder	(**classifier**/transformer/rnn)
batch size	(1000/2000/**3000**)
train steps	10,000
dropout	0.1
learning rate	$2e^{-3} \cdot min\left(step^{-0.5}, step \cdot warmup^{-1.5}\right)$
warmup	(**1000**/10,000)
optimizer	adam

Table 2 Hardware and software environment

Device	Congiguration
Operating system	Ubuntu 20.04.6 LTS
Processor	Intel Xeon(R) Gold6133 CPU ®2.50GHz
GPU	RTX 3090 (24GB)*2
Framework	Pytorch
Compilers	PyCharm
Scripting language	Python 3.8

Table 4 Hyper-parameters of BART

Parameters	Values
batch_size	32
epochs	3
max_input_length	521
max_target_length	150
learning_rate	1e-04
warmup_steps	10
weight_decay	0.001
metric_for_best_model	ROUGE-1

Chinese prompt	English prompt
1.根据医疗对话{***ls_dialog***}生成一段简短的医学摘要. 2.包含以下六个部分：	1.Generate a short medical summary based on the medical dialogue{***ls_dialog***}. 2.Contains the following six parts:
主诉:	**Chief complaint**:
现病史:	**Present medical history**:
辅助检查:医学检查项目名称	**Auxiliary examination**: Medical examination names
既往史:	**Past medical history**:
诊断:最相关的疾病名称	**Diagnostics**: The most relevant diagnostic name
建议:	**Recommendation**:

Fig. 8 A simple prompt for medical dialogue summarization without any complex parameter variables, abbreviated as Prompt_S

Chinese prompt	English prompt
根据医患对话{*ls_dialog*}生成一段简短的医学摘要，包含以下六个部分：	Generate a short medical summary based on the doctor-patient conversation {*ls_dialog*} with the following six parts:
主诉:不超过20字，包括以症状{*ls_symptom*}为例，根据医患对话生成症状，症状天数。	**Chief complaint**: Not exceeding 20 words, including the symptom {*ls_symptom*} as an example, according to the doctor-patient dialogue to generate symptoms, symptoms of the number of days.
现病史:不超过80字，生成相关的现病史，以{*ls_pre_history*}为例。	**Present medical history**: Not exceeding 80 words, generate related history of disease, {*ls_pre_history*} as an example.
辅助检查:生成最相关的医学检查名称，以{*ls_aux_test*}为例。	**Auxiliary examination**: Generate the most relevant medical examination name, using {*ls_aux_test*} as an example.
既往史:不超过50字，生成相关的既往史，以{*ls_past_history*}为例。	**Past medical history**: Not exceeding 50 characters to generate the relevant Past medical history, for example, {*ls_past_history*}.
诊断:生成一个最相关的诊断名称，以{*ls_diagnosis*}为例。	**Diagnostics**: Generates the most relevant diagnosis name, for example, {*ls_diagnosis*}.
建议:不超过80字,生成所有建议。	**Recommendation**: Not exceeding 80 words, generate all recommendations.

Fig. 9 A technical prompt for medical dialogue summarization with some parameter variables, abbreviated as prompt_T

the corresponding contents in Fig. 10, which will not be repeated here.

Evaluation metrics

Automatic evaluation metrics

In this paper, the ROUGE-1, ROUGE-2, ROUGE-L [50] and BERTScore [51] as automatic evaluation metrics.

Given a reference summary $x = (x_1, \ldots, x_k)$ and a candidate summary $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_k)$. By embedding model generate reference summary vector $X = (X_1, \ldots, X_m)$ and candidate summary vector $\hat{X} = \left(\hat{X}_1, \ldots, \hat{X}_m\right)$, respectively.

ROUGE-n represents an n-gram recall measure comparing a reference summary to the corresponding candidate summary. The computation of ROUGE-n is as follows:

$$P_{ROUGE-n} = \frac{Count_{match}\left(gram_n \in \left(x, \hat{x}\right)\right)}{Count\left(gram_n \in \hat{x}\right)}$$
$$R_{ROUGE-n} = \frac{Count_{match}\left(gram_n \in \left(x, \hat{x}\right)\right)}{Count\left(gram_n \in x\right)} \quad (9)$$
$$F_{ROUGE-n} = \frac{2 * P_{ROUGE-n} * R_{ROUGE-n}}{P_{ROUGE-n} + R_{ROUGE-n}}$$

Where n denotes the length of the n-gram, so $gram_n$ represents the 1-gram or 2-gram, and $Count_{match}\left(gram_n \in \left(x, \hat{x}\right)\right)$ signifies the maximum number of 1-gram or 2-gram co-occurring in a reference summary and the corresponding candidate summary.

Variable-Annotation	Chinese	English
{ls_dialog} A multi-turn doctor-patient dialog	患者:宝宝7个月轮状病毒腹泻已经1个月了,吃过双岐妈咪爱蒙脱石散都不管用,一天拉 4.5次,今天有点发烧不到 38 度该怎么处理。医生：你好。患者：您好。医生：目前这些药物服用多久了？一点效果都没有吗？患者：一直服用就是有时会少服一些，没有定量服用。医生：有大便常规化验单吗，发我看下吧。患者：好的。医生：好。	Patient: The baby is 7 months old rotavirus diarrhea has been 1 month, eat Shuangqi mommy love montmorillonite powder is no use, pull 4.5 times a day, today a little fever less than 38 degrees how to deal with. Doctor: Hello patient: hello. Doctor: How long have these drugs been taken? It didn't work at all? Patient: I have been taking it all the time, but sometimes I will take some less. Doctor: Do you have a routine stool test? Please send it to me. Patient: good. Doctor: OK.
{ls_symptom} Related symptoms in children	发热、咳嗽、喉咙痛、流鼻涕、腹痛、腹泻、喷嚏、呕吐、拉肚子、肚子疼、疲倦、乏力等。	Fever, cough, sore throat, runny nose, abdominal pain, diarrhea, sneezing, vomiting, loose stools, stomach pain, fatigue, fatigue, etc.
{ls_pre_history} Present medical history	当前的症状、发生部位、发生时间、频率、持续时间、严重程度、伴随症状、药物治疗等。	Current symptoms, location, time of occurrence, frequency, duration, severity, concomitant symptoms, medication, etc.
{ls_aux_test} Auxiliary examination	血常规、尿常规、粪便常规、支原体、肝功能、肾功能、凝血功能、血型鉴定、X射线、CT扫描、MRI、超声波等。	Complete blood count (cbc), urinalysis, stool routine, chlamydia test, liver function test (lft), kidney function test (kft), coagulation profile, blood typing, x-ray, ct Scan, mri, ultrasound, etc.
{ls_past_history} Past medical history	既往病史：患者曾经患过的疾病，如高血压、糖尿病、哮喘等。手术史：患者曾经接受过的手术，包括手术的种类、时间和结果等。药物过敏史：患者对某些药物或物质存在过敏反应的情况，如青霉素过敏等。家族病史：患者的家族成员是否有某些疾病的遗传倾向或患病情况，如高血压、糖尿病、心脏病等。	Previous medical history: Diseases that the patient had ever had, such as high blood pressure, diabetes, asthma, etc. Surgical history: The surgery the patient has undergone, including the type, timing, and outcome of the surgery. Drug allergy history: A condition in which the patient has an allergic reaction to certain drugs or substances, such as penicillin allergy. Family history: Whether the patient's family members have a genetic predisposition to certain diseases or conditions, such as high blood pressure, diabetes, heart disease, etc.
{ls_diagnosis} Diagnostics	小儿支气管炎、小儿发热、小儿腹泻、上呼吸道感染、小儿消化不良、小儿感冒、小儿咳嗽、新生儿黄疸、小儿便秘、小儿支气管肺炎等。	Pediatric bronchitis, pediatric fever, pediatric diarrhea, upper respiratory infection, pediatric indigestion, pediatric common cold, pediatric cough,neonatal jaundice, pediatric constipation, pediatric bronchopneumonia, etc.

Fig. 10 Detailed description of the parameters related to prompt_T

We known that the longer the Longest Common Subsequence (LCS) of two summary sentences, the more similar the two summaries are. We employ the ROUGE-L to measure the similarity between a reference summary and the corresponding candidate summary. The calculation is as follows:

$$
\begin{aligned}
P_{ROUGE-L} &= \frac{LCS\left(x,\hat{x}\right)}{\left|\hat{x}\right|}\\
R_{ROUGE-L} &= \frac{LCS\left(x,\hat{x}\right)}{|x|}\\
F_{ROUGE-L} &= \frac{2*P_{ROUGE-L}*R_{ROUGE-L}}{P_{ROUGE-L}+R_{ROUGE-L}}
\end{aligned}
\tag{10}
$$

Where $LCS\left(x,\hat{x}\right)$ is the length of a longest common subsequence of the reference summary and the corresponding candidate summary. $|x|$ is the length of the reference summary and $\left|\hat{x}\right|$ is the length of the candidate summary.

BERTScore computes the complete score by matching each token in candidate summary $\hat{x}$ to a token in reference summary x to calculate precision, and each token in reference summary x to a token in candidate summary $\hat{x}$ to calculate recall .

$$
\begin{aligned}
COS\left(X_i,\hat{X}_j\right) &= \frac{X_i^T\hat{X}_j}{\|X_i\|\left\|\hat{X}_j\right\|}\\
P_{BERT} &= \frac{1}{\left|\hat{x}\right|}\sum_{\hat{x}_j\in\hat{x}}\max_{x_i\in x}\left(COS\left(X_i,\hat{X}_j\right)\right)\\
R_{BERT} &= \frac{1}{|x|}\sum_{x_i\in x}\max_{\hat{x}_j\in\hat{x}}\left(COS\left(X_i,\hat{X}_j\right)\right)\\
F_{BERT} &= \frac{2*P_{BERT}*R_{BERT}}{P_{BERT}+R_{BERT}}
\end{aligned}
\tag{11}
$$

Where $COS\left(X_i,\hat{X}_j\right)$ is the cosine similarity of a reference summary x and a candidate summary $\hat{x}$ is given by the formula $\frac{X_i^T\hat{X}_j}{\|X_i\|\left\|\hat{X}_j\right\|}$. $|x|$ is the length of the reference summary and $\left|\hat{x}\right|$ is the length of the candidate summary.

Human evaluation metrics

We recruited three domain experts with medical training, and each of them individually annotated 100 randomly selected medical dialog samples from the dataset. In total, we collected 300 annotations, with three annotations for each sample. **Contains Key Result**, **Coherence**, **Usefulness** and **Readability** as human evaluation metrics [48].

Results

In Table 5, "Prompt_S" and "Prompt_T" represent the use of simple and technical prompt modes in ChatGPT. The combinations of "Temperature" and "Top_p" parameters for these two different prompts can be adjusted to modify the generation results of the ChatGPT model, catering to the requirements of medical dialogue summaries. The "Temperature" parameter is used to control the randomness and creativity of generated text, while the "Top_p" parameter is used to control the diversity of generated text. A higher "Top_p" value results in more diverse text, and a lower "Top_p" value results in more consistent text. Higher "Temperature" and higher "Top_p" values create more randomness and creativity, but may result in generated content that is less relevant to the input.

From Fig. 11, it can be seen that when Temperature=1.0 and Top_p=1.0, ChatGPT generated a medication recommendation as the diagnosis result. When Temperature=0.7 and Top_p=1.0, ChatGPT generated a medication name as the diagnosis result.

Although this situation is relatively rare and doesn't happen every time, it indirectly confirms that excessively high Temperature and Top_p parameter values may have a negative impact on the summary results. Conversely, lower "Temperature" and lower "Top_p" values make the generated content more conservative and relevant, but potentially less innovative. These parameters can be adjusted according to the actual situation of the specific application and requirements. As shown in Fig. 12, when parameters "Temperature" and "Top_p" are set between 0.1 and 1, the variation trend of the ROUGE score is basically consistent with the characteristics of parameters "Temperature" and "Top_p" themselves, especially with the decrease of the value of the "Top_p" parameter, the ROUGE-1 and ROUGE-L scores increase.

The results from Table 6 indicate that when ChatGPT is in modes "Prompt_S" and "Prompt_T", which correspond to "Temperature=1.0" and "Top_p=1.0" the ROUGE-1 and ROUGE-L scores are the lowest. When "Temperature=0.1" and "Top_p=0.1", the ROUGE-1 and ROUGE-L scores are the highest. This indicates that adjusting the "Temperature" and "Top_p" parameters appropriately based on their characteristics can indeed influence the final ROUGE results. Furthermore, using "Prompt_T" as ChatGPT's prompt yields significantly better results compared to using "Prompt_S". This indicates that well-designed prompts can significantly enhance the performance of ChatGPT in generating summaries.

We randomly selected 100 sets of medical dialogue summaries generated by the BERTSUM, BART, and ChatGPT models, where ChatGPT used prompts Prompt_S.7 and Prompt_T.7 to generate the medical dialogue summaries. Nevertheless, the pre-trained model approach based on BART outperformed the best results obtained with ChatGPT in both prompt modes, and was 14.94% better than the highest value corresponding to ChatGPT on the ROUGE-1 score and 32.84% better than the highest score corresponding to ChatGPT on the ROUGE-L score. From Table 7, it is evident that the dialogue summaries generated by

Table 5 Temperature and Top_p parameter combinations for ChatGPT's prompt model, such as Prompt_S and Prompt_T

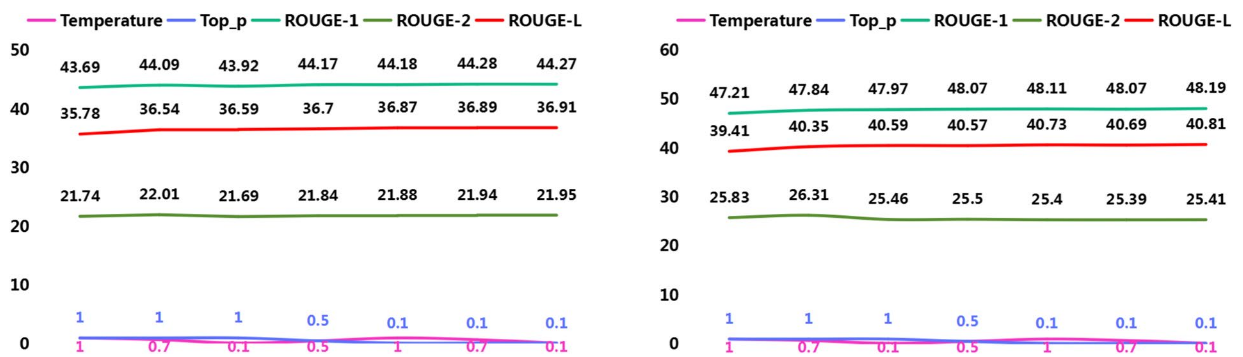
Prompt Model	Temperature	Top_p
Prompt_S.1	1.0	1.0
Prompt_S.2	0.7	1.0
Prompt_S.3	0.1	1.0
Prompt_S.4	0.5	0.5
Prompt_S.5	1.0	0.1
Prompt_S.6	0.7	0.1
Prompt_S.7	0.1	0.1
Prompt_T.1	1.0	1.0
Prompt_T.2	0.7	1.0
Prompt_T.3	0.1	1.0
Prompt_T.4	0.5	0.5
Prompt_T.5	1.0	0.1
Prompt_T.6	0.7	0.1
Prompt_T.7	0.1	0.1

Table 6 Rouge and BERTScore scores for ChatGPT's prompt model, such as Prompt_S and Prompt_T

Prompt Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Prompt_S.1	43.69	21.74	35.78	71.43
Prompt_S.2	44.09	22.01	36.54	71.47
Prompt_S.3	43.92	21.69	36.59	71.31
Prompt_S.4	44.17	21.84	36.67	71.42
Prompt_S.5	44.18	21.88	36.87	71.40
Prompt_S.6	44.28	21.94	36.89	71.48
Prompt_S.7	44.27	21.95	36.91	71.49
Prompt_T.1	47.21	25.83	39.41	73.34
Prompt_T.2	47.84	**26.31**	40.35	73.38
Prompt_T.3	47.97	25.46	40.59	73.32
Prompt_T.4	48.07	25.50	40.57	73.36
Prompt_T.5	48.11	25.40	40.73	73.37
Prompt_T.6	48.07	25.39	40.69	73.37
Prompt_T.7	**48.19**	25.41	**40.81**	**73.38**

Type	Chinese	English
Manual summary	诊断：消化不良。	Diagnosis: Indigestion.
ChatGPT summary (Temperature = 1.0 Top_P = 1.0)	诊断: 小儿消积止咳口服液可以缓解症状。	Diagnosis: Pediatric phlegm-reducing and cough-relieving oral liquid can alleviate the symptoms.
ChatGPT summary (Temperature = 0.7 Top_P = 1.0)	诊断：可能为小儿消积止咳口服液。	Diagnosis: It may be pediatric phlegm-reducing and cough-relieving oral liquid

Fig. 11 Higher values of the Temperature and Top_P parameters may lead to partial summary results generated by ChatGPT that may be inconsistent with the actual situation



(a) Human evaluation metrics of ChatGPT model based on Prompt_S.

(b) Human evaluation metrics of ChatGPT model based on Prompt_T.

Fig. 12 The "Temperature" parameter controls the level of randomness and creativity in the generated text, while the "Top_p" parameter influences the diversity of the generated content. A higher "Top_p" value leads to more diverse text, whereas a lower value results in more consistent text. Elevated values for both "Temperature" and "Top_p" introduce greater randomness and creativity but may reduce the relevance of the generated content to the input. Conversely, lower values for both parameters make the generated content more conservative and relevant but potentially less innovative

Table 7 Automatic evaluation metrics for BERTSUM, BART, and ChatGPT summaries, such as comparisons of ROUGE-1, ROUGE-2, ROUGE-L and BERTScore scores

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
BERTSUM_Classifier	34.51	13.94	24.47	63.53
BERTSUM_Transfromer	33.52	13.21	24.06	63.21
BERTSUM_RNN	33.73	13.62	24.17	63.18
BART	**55.39**	**40.38**	**54.21**	**78.32**
ChatGPT(Prompt_T.7)	48.19	25.41	40.81	73.38

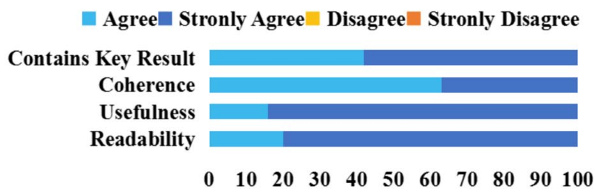
the BART model outperform BERTSUM and ChatGPT in terms of automatic evaluation metrics, with higher ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore scores. However, the effect of ChatGPT is much better than BERTSUM and BART models under the human evaluation metrics.

Discussion

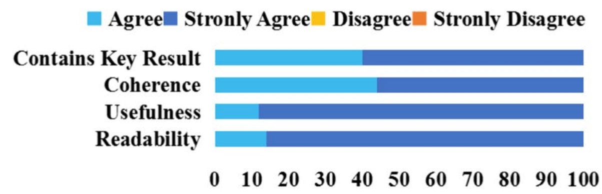
Comparison between automatic evaluation metrics and human evaluation metrics

As depicted in sub-figures (a) and (b) of Fig. 13, medical experts showed high approval of the summary quality generated by ChatGPT, with higher approval rates in Prompt_T compared to Prompt_S. As depicted in sub-figure (c) of Fig. 13. Except for the "Readability" metric, the other three metrics had barely passed the medical experts' evaluation, with less than 30% of the summaries meeting the criteria. In contrast, the summaries generated by ChatGPT completely passed the medical experts' evaluation, with many of them being rated as "Strongly Agree" as shown in sub-figures (a) and (b) of Fig. 13. Sub-figure (d) in Fig. 13 illustrates that medical experts have rated the medical dialogue summaries generated by the BERTSUM model as mostly "Strongly Disagree" on almost all human evaluation metrics. This indicates that such summaries have no reference value for users.

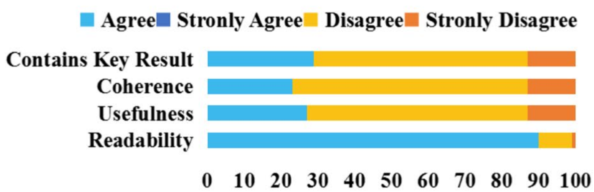
Additionally, from Figs. 14 and 15, we can observe that the summaries generated by the BART model are



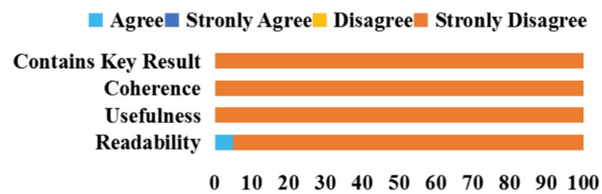
(a) Human evaluation metrics of ChatGPT model based on Prompt_S.



(b) Human evaluation metrics of ChatGPT model based on Prompt_T.



(c) Human evaluation metrics of BART.



(d) Human evaluation metrics of BERTSUM.

Fig. 13 Human evaluation of 100 summaries generated by ChatGPT, BART, and BERTSUM models, with average scores on four evaluation metrics: Contains Key Result, Coherence, Usefulness, and Readability. Sub-figures (**a**) and (**b**) demonstrate that summaries generated by ChatGPT achieved favorable results in the human evaluation metrics, especially under the Prompt_T condition, with a substantial proportion of "Strongly Agree" in all metrics. However, sub-figure (**c**) indicates that the BART model performed poorly in the human evaluation metrics, except for the "Readability" metric. sub-figure (**d**) shows that the BERTSUM model exhibited very poor performance across all metrics, almost entirely in the "Strongly Disagree" state

relatively short. After conducting statistical analysis, it was found that the manually annotated summaries in the original dataset had an average length of 100 words, while the average length of summaries generated by the BART model was only 60 words. This discrepancy could possibly be attributed to the influence of the style of manual annotations. Although the generated summaries may appear concise, they tend to overlook crucial information, leading to a reduction in their overall informative value. In contrast, ChatGPT generates summaries of approximately 200 words, which are perceived by human experts as more comprehensive, effective, and valuable.

A few real-world examples of BART and ChatGPT models on medical dialog summaries

As shown in Fig. 16, the "Recommendation" part of the manual summary lists "aluminum magnesium carbonate tablets" alongside "probiotics" and "montmorillonite powder for children" as medications, without clearly specifying whether "aluminum magnesium carbonate tablets" is intended for the patient's child or the patient themselves, this lack of clarity could lead to misunderstanding. On the other hand, the "Recommendation" part of ChatGPT's summary clearly states, "patients can use aluminum magnesium carbonate tablets to neutralize stomach acid", indicating that ChatGPT can significantly determine that the medication is for the patient themselves, not their child. Additionally, in Fig. 17, the "Recommendation" part of ChatGPT's summary includes advice such as "to maintain the baby's water intake, you can give an appropriate amount of oral rehydration salt solution", which was not explicitly mentioned in the original conversation. However, medical experts recognize that ChatGPT's generated advice is entirely consistent with medical knowledge, given the context of the original conversation where it mentions concern about the child's dehydration due to frequent diarrhea after feeding. In this context, ChatGPT provides more reasonable recommendations than the manual summary. In Fig. 14, the "Diagnosis" output from the BART summary is "Upper respiratory infection" while the correct diagnosis should be related to a disease associated with "Diarrhea". This incorrect diagnosis is a significant discrepancy in the BART summary. Additionally, the BART summary is overly concise, leading to the omission of some potentially important information. In the "Recommendation" part, BART's summary only mentions the recommendation of "Oral montmorillonite powder". However, the original conversation actually includes additional recommendations such as "routine stool examination and other relevant examination" and "avoid eating greasy, spicy and irritating food,and feed more liquid food". The exclusion of these important recommendations in the BART summary results in the loss of crucial information and diminishes the overall usefulness of the summary. In Fig. 15, the BART summary closely resembles the manual

Type	Chinese	English
Manual summary	**主诉**：腹泻 2 天。 **现病史**：患儿出现腹泻2天，伴呕吐，去当地医院诊疗,肌肉针注射及服用蒙脱石散治疗。 **辅助检查**：无。 **既往史**：无。 **诊断**：腹泻。 **建议**：蒙脱石散加量。	**Chief complaint**: Diarrhea for 2 days. **Present medical history**: The child developed diarrhea for 2 days with vomiting, went to a local hospital for diagnosis and treatment, muscle needle injection and montmorillonite powder treatment. **Auxiliary examination**: None. **Past medical history**:None. **Diagnosis**: Diarrhea. **Recommendation**: Dosage of montmorillonite powder.
BART summary ROUGE-1 = 60.00	**主诉**：腹泻。 **现病史**：患儿腹泻，大便。 **辅助检查**：暂缺。 **既往史**：不详。 **诊断**：上呼吸道感染。 **建议**：口服蒙脱石散。	**Chief complaint**: Diarrhea. **Present medical history**: Diarrhea, stool. **Auxiliary examination**: Temporarily missing. **Past medical history**: Unknown. **Diagnosis**: Upper respiratory infection. **Recommendation**: Oral montmorillonite powder.
ChatGPT summary ROUGE-1 = 32.31	**主诉**：宝宝八个月，拉肚子俩天了。 **现病史**：宝宝喝奶粉,最近没有更换奶粉,也没有吃不干净的东西，但吃了点自己做的汤圆(糯米粉)。家里没有人腹泻。起初宝宝感冒，喂奶粉后出现呕吐，接着下午开始拉肚子。到附近的卫生乡镇医院打了小针，但没有告知是什么针。由于是农村，还没有到省里医院进行检查。 **辅助检查**：建议进行大便常规检查和拍摄大便照片。 **既往史**：没有提到相关既往史。 **诊断**：根据症状，可初步诊断为急性腹泻。 **建议**：1.带宝宝到省里医院进行进一步检查,包括大便常规检查和其他相关检查。2. 在家注意给宝宝提供清淡易消化的食物，避免食用油腻、辛辣刺激性食物，多喂食流质食物。3. 给宝宝补充足够的水分，避免脱水。4.按医生嘱咐正确使用蒙脱石散液体,加量使用,提高疗效。	**Chief complaint**: The baby is eight months old and has diarrhea for two days. **Present medical history**: The baby drinks milk powder, has not changed the milk powder recently, and has not eaten anything unclean, but has eaten some tangyuan (glutinous rice flour) made by himself. No one in the family had diarrhea. At first, the baby caught a cold, vomited after feeding milk powder, and then began to diarrhea in the afternoon. He went to a nearby health township hospital to get a small injection, but was not told what the injection was. Because it is a rural area, I have not been to the provincial hospital for examination. **Auxiliary examination**: It is recommended to perform routine stool examination and take stool photos. **Past medical history**: No relevant past medical history is mentioned. **Diagnosis**: Based on the symptoms, the initial diagnosis can be acute diarrhea. **Recommendation**: 1. Take the baby to the provincial hospital for further examination,including routine stool examination and other relevant examinations. 2. At home, pay attention to provide the baby with light and easy to digest food, avoid eating greasy, spicy and irritating food, and feed more liquid food. 3. Give your baby enough water to avoid dehydration. 4. Use montmorillonite powder correctly according to the doctor's instructions, and increase the amount to improve the curative effect.

Fig. 14 From the perspective of ROUGE-1 score, the BART summary here shows a high similarity to the manual summary. However, there are significant issues with the BART summary. Firstly, in the "Diagnosis" part, the BART summary incorrectly states the diagnosis as "Upper respiratory infection", while the correct diagnosis in the manual summary is "Diarrhea". Secondly, the entire summary is too brief, leading to the omission of some potentially important information. For instance, in the "Recommendation" part, the BART summary only mentions the recommendation of "Oral montmorillonite powder". Although ChartGPT's ROUGE-1 score is lower than BART's, the resulting summary is highly detailed and semantically consistent with the original conversation data, such as "routine stool examination and other relevant examinations" and "avoid eating greasy, spicy and irritating food, and feed more liquid food"

summary, and its ROUGE-1 score is also high. However, in terms of practical effectiveness, especially in the "Recommendation" part where the content is "Continue to take oral medications for cold medicine", such content provides a rather vague recommendation and lacks useful information. On the other hand, the ChatGPT summary offers more detailed advice. In addition to recommending the medication "spleen ammonia peptide freeze-dried powder", it also suggests "atomization" and "make some pear tea for the baby to drink". In Fig. 18, the main issues with the summaries generated by ChatGPT are: (1). The "Chief complaint" part is overly lengthy. (2). In the "Auxiliary examination" part, there are suggestions for examinations that did not actually occur. However, despite these issues, they do not affect the understanding of the generated summaries by both the medical professionals and patients.

In summary, these examples illustrate that ChatGPT can significantly enhance the accuracy and specificity of recommendations by considering contextual information and generating more appropriate advice than manual summaries. Moreover, ChatGPT proves to be more effective and valuable in generating summaries compared to BERTSUM and BART. However, this also serves as a reminder that when assessing the quality of summaries, it is essential to consider not only automatic evaluation metrics such as ROUGE but also conduct comprehensive analyses, taking into account the actual content and application scenarios. By taking a holistic approach to evaluation, we can better understand the capabilities and

Type	Chinese	English
Manual summary	**主诉：**咳嗽。 **现病史：**今早咳嗽。 **辅助检查：**暂缺。 **既往史：**不详。 **诊断：**上呼吸道感染。 **建议：**观察多喝水，服用治疗感冒的中成药，脾氨肽冻干粉。	**Chief complaint**: Cough. **Present medical history**: I coughed this morning. **Auxiliary examination**: Temporarily missing. **Past medical history**:Unknown. **Diagnosis**: Upper respiratory infection. **Recommendation**: Observe to drink more water, take the treatment of cold Chinese patent medicine, spleen ammonia peptide freeze-dried powder.
BART summary ROUGE-1 = 73.91	**主诉：**咳嗽。 **现病史：**患儿咳嗽。 **辅助检查：**暂缺。 **既往史：**不详。 **诊断：**上呼吸道感染。 **建议：**继续口服药物治疗感冒药。	**Chief complaint**: Cough. **Present medical history**: Patient with cough. **Auxiliary examination**: Temporarily missing. **Past medical history**: Unknown. **Diagnosis**: Upper respiratory infection. **Recommendation**: Continue to take oral medications for cold medicines.
ChatGPT summary ROUGE-1 = 56.44	**主诉：**宝宝咳嗽，每到换季容易感冒。 **现病史：**宝宝今天早上起来开始咳嗽，没有发热、流鼻涕、打喷嚏等症状。考虑为上呼吸道感染。 **辅助检查：**无需辅助检查。 **既往史：**无特殊既往史。 **诊断：**上呼吸道感染。 **建议：**可以吃治疗感冒的中成药，观察体温情况，多喝水。严重的话可以做雾化。可以煮些梨子茶给宝宝喝。可以考虑吃调节免疫的药,如脾氨肽冻干粉。注意梨水不要喝太多,以免引起拉肚子。	**Chief complaint**: The baby coughs, easy to catch a cold every season. **Present medical history**: The baby began to cough this morning, no fever, runny nose, sneezing and other symptoms. An upper respiratory infection was considered. **Auxiliary examination**: No auxiliary examination is required. **Past medical history**: No special Past medical history. **Diagnosis**: Upper respiratory infection. **Recommendation**: You can eat the treatment of cold Chinese patent medicine, observe the temperature, drink more water. In severe cases, atomization can be done. You can make some pear tea for the baby to drink. Consider taking immune-regulating drugs, such as spleen ammonia peptide freeze-dried powder. Be careful not to drink too much pear water, so as not to cause diarrhea.

Fig. 15 From the perspective of ROUGE-1 score, the summary generated by BART is highly similar to the manual summary. However, in terms of practical effectiveness, especially in the "Recommendation" part where the content is "Continue to take oral medications for cold medicines", such content provides a rather vague recommendation and lacks useful information. On the other hand, the advice given by ChatGPT is more detailed and valuable. For instance, in addition to recommending the medication "spleen ammonia peptide freeze-dried powder", it also suggests "atomization" and "make some pear tea for the baby to drink". Such specific and practical information can offer more assistance and guidance to the readers

Type	Chinese	English
Manual summary	**主诉**: 大便异常。 **现病史**: 患儿果冻样绿色大便，每天五六次。 **辅助检查**: 暂缺。 **既往史**: 不详。 **诊断**: 腹泻。 **建议**: 益生菌，小儿蒙脱石散，铝碳酸镁片。	**Chief complaint**: Abnormal stool. **Present medical history**: The child had jelly-like green stool five or six times a day. **Auxiliary examination**: Temporarily missing. **Past medical history**:Unknown. **Diagnosis**: Diarrhea. **Recommendation**: Probiotics, children montmorillonite powder, aluminum magnesium carbonate tablets.
ChatGPT summary	**主诉**：小孩拉绿色大便，五六次/天，持续多天。 **现病史**：小孩拉绿色大便已持续多天，每天五六次，伴有频繁更换尿片。小孩一天泡一两次奶粉，同时也有母乳喂养。患者自身感冒，担心是否会影响小孩。 **辅助检查**：无。 **既往史**： 无。 **诊断**：小儿腹泻。 **建议**：使用小儿蒙脱石散和益生菌调理，选择四联益生菌效果更好。同时，患者可以使用铝碳酸镁片中和胃酸。注意保暖，避免小孩受凉。	**Chief complaint**: The child pulled green stool, five or six times/day, lasting for several days. **Present medical history**: The child has been pulling green stool for several days, five or six times a day, accompanied by frequent diaper changes. The child was given milk powder once or twice a day and breastfed. The patient himself has a cold and is worried about whether it will affect the child. **Auxiliary examination**: None. **Past medical history**: None. **Diagnosis**: Pediatric diarrhea. **Recommendation**: Use children montmorillonite powder and probiotic conditioning, choose four probiotics effect is better. At the same time, patients can use aluminum magnesium carbonate tablets to neutralize stomach acid. Keep warm and keep children from getting cold.

Fig. 16 The manual summary mistakenly leads people to believe that "aluminum magnesium carbonate tablets" is intended for children, but in reality, it is meant for parents of children. On the other hand, ChatGPTis able to distinguish between different patients in the context of the conversation, i.e. the user of the drug is clearly distinguished by "children" and "patient", where "children" means the sick child and "patient" means the parents of the sick child. For example, "patients can use aluminum magnesium carbonate tablets to neutralize stomach acid"

Type	Chinese	English
Manual summary	**主诉**：腹泻伴发烧 2 天。 **现病史**： 2天前患儿无明显诱因下出现腹泻伴发烧症状，粪便呈稀水样便，发烧体温最高时达 38.7℃。1 天前复测体温达 37.7℃。无咳嗽咳痰,无恶心呕吐,无其他明显不适症状。精神状态一般,胃纳一般,余如常。 **辅助检查**：暂缺。 **既往史**：暂缺。 **诊断**：小儿腹泻。 **建议**：忌强迫进食,密切观察,必要时及时就医,查便常规等检查明确相关病因后,对因治疗。	**Chief complaint**: Diarrhea with fever for 2 days. **Present medical history**: 2 days ago, the child developed diarrhea with fever without obvious cause. The stool was watery and the highest temperature of fever reached 38.7℃. 1 day ago, the retest temperature reached 37.7℃. No cough, no phlegm, no nausea, no other obvious symptoms of discomfort. Normal mental state, normal stomach, normal. **Auxiliary examination**: Temporarily missing. **Past medical history**:Not available. **Diagnosis**: Pediatric diarrhea. **Recommendation**: Avoid forced eating, close observation, timely medical treatment if necessary, routine checks and other tests to clear the cause of the cause of treatment.
ChatGPT summary	**主诉**：宝宝十个月,腹泻发烧三天。 **现病史**：宝宝前天开始腹泻发烧,前天体温38.7 度，拉 2 次，较稀，昨天体温 37.7 度，拉一次,水状便,今天依然低烧,每次吃奶后都拉,水样便,不爱吃奶和辅食。 辅助检查：无。 **既往史**：无。 **诊断**：小儿腹泻。 **建议**：您可以观察宝宝的体温变化和大便情况,如果宝宝体温持续升高或出现呕吐、脱水等症状,建议及时就医。同时,保持宝宝的水分摄入量,可以适量给予口服补液盐溶液。如果宝宝的症状持续加重或不缓解，建议尽快就医进行进一步检查和治疗。	**Chief complaint**: The baby was ten months old and had diarrhea and fever for three days. **Present medical history**: The baby began to have diarrhea and fever the day before yesterday, the temperature of the day before yesterday was 38.7 degrees, pulled twice, relatively thin, yesterday's temperature was 37.7 degrees, pulled once, watery stool, today is still low fever, pulled every time after eating milk, watery stool, do not like to eat milk and complementary foods. **Auxiliary examination**: None. **Past medical history**: None. **Diagnosis**: Pediatric diarrhea. **Recommendation**: You can observe the baby's temperature changes and stool, if the baby's temperature continues to rise or vomiting, dehydration and other symptoms, it is recommended to seek medical attention in time. At the same time, to maintain the baby's water intake, you can give an appropriate amount of oral rehydration salt solution. If your baby's symptoms continue to worsen or do not ease, it is recommended to seek medical attention as soon as possible for further examination and treatment.

Fig. 17 In the original conversation, it is evident that the child is suffering from diarrhea with watery stools. Generally, in such cases, doctors would recommend oral rehydration with a saline solution to the patient to prevent dehydration. However, this advice is not evident from the manual summary, primarily because the original text did not mention the relevant content of oral rehydration. On the other hand, ChatGPT can directly provide a reasonable recommendation.Such as "to maintain the baby's water intake, you can give an appropriate amount of oral rehydration salt solution"

limitations of language models such as ChatGPT, and ensure that their outputs align with real-world use cases and human expectations. ChatGPT's ability to provide contextually relevant and useful recommendations highlights its potential in various natural language processing tasks, and it emphasizes the importance of responsible evaluation practices in the development and deployment of AI systems.

Conclusion

The study compares the performance of the BART, ChatGPT, and BERTSUM models in generating medical dialogue summaries. The results indicate that summaries generated by the BERTSUM model exhibit notably lower ROUGE and BERTScore scores, and fail human evaluation across all metrics. Conversely, the BART model achieves the highest ROUGE and BERTScore scores, outperforming ChatGPT. It is ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore scores surpass ChatGPT's best results by 14.94%, 53.48%, 32.84%, and 6.73% respectively. However, in human evaluation by medical experts, BART's summaries perform well only in "Readability" with less than 30% passing evaluation in other metrics. Compared to BERTSUM and BART, the ChatGPT model is preferred by human medical experts. In conclusion, ChatGPT can manipulate medical dialogue summary style and outcomes using various prompts. The generated content is not only better received than certain human experts' results but also more comprehensible, showing promise for automated medical dialogue summarization. However, automatic evaluation metrics such as ROUGE and BERTScore may have limitations when it comes to comprehensively assessing the outputs of large language models like ChatGPT, therefore, further research is needed to explore more suitable evaluation metrics. Additionally, there are still some issues with the medical dialogue summaries generated by ChatGPT, such as overly lengthy "Chief Complaint" part and the inclusion of certain tests in the "Auxiliary examination" part that did not actually occur, and one more, improperly configured fine-tuning parameters for ChatGPT can indeed lead to incorrect results. In conclusion, ChatGPT's performance in medical conversation summarization is influenced by various factors. Therefore, future research needs to further identify the key factors affecting model output results and solve them systematically step by step.

Type	Chinese	English
Manual summary	**主诉**：腹泻1周,伴泡沫样绿色粪便。 **现病史**：1周前患儿无明显诱因下出现腹泻,伴泡沫样绿色粪便,偶见白色粘液脓样物质。无咳嗽咳痰，无其他明显不适症状。监护人予思密达自行服用,未见明显好转。后于当地医院就医后予蒙脱石散及乳酸菌服用2天,具体检查及结果不详。精神状态一般,胃纳一般,余如常。 **辅助检查**：暂缺。 **既往史**：暂缺。 **诊断**：小儿腹泻。 **建议**：及时就医查粪便常规明确病因后对因治疗，暂予妈咪爱、双歧杆菌服用，密切观察。	**Chief complaint**: Diarrhea for 1 week, accompanied by frothy green stools. **Present medical history**: The child developed diarrhea without apparent cause 1 week ago, accompanied by frothy green stools with occasional presence of white mucopurulent material. No cough or phlegm, and no other notable discomfort symptoms. The guardian administered Simethicone for the child, but there was no significant improvement. The child was subsequently taken to a local hospital, where they were given Attapulgite Powder and lactobacillus for 2 days. Specific examination details and results are not available. The child's mental state is normal, appetite is normal, and other aspects are as usual. **Auxiliary examination**: Temporarily missing. **Past medical history**:Temporarily missing. **Diagnosis**: Pediatric diarrhea. **Recommendation**: Seek prompt medical attention for fecal routine examination to identify the cause and provide specific treatment. Administer Mama Love and probiotics for now, and closely monitor the child's condition.
ChatGPT summary ROUGE-1 = 47.55	**主诉：**孩子一周，一直拉屎，屎的症状有点消化不良，有的时候呈绿色，有的时候是沫沫，有的时候里面掺加白色的粘东西,吃了两天思密达,后来去医院看,吃了两天蒙脱石散和乳杆菌,还是不见好,需不需要吃点消炎药，吃什么。 **现病史**：孩子一周来一直有腹泻，症状表现为拉屎颜色不正常，有时呈绿色，有时是沫沫，有时还有白色的粘东西。孩子之前吃了思密达、蒙脱石散和乳杆菌，但症状没有改善。 **辅助检查**：建议去医院查大便，以排除感染性肠炎和进一步确定病因。 **既往史**：无相关既往史。 **诊断**：可能为消化不良引起的腹泻。 **建议**：建议去医院查大便以确定病因,并根据检查结果进行治疗。在等待检查结果期间,继续服用益生菌和蒙脱石散。避免盲目使用消炎药物,以免对消化道造成伤害。注意观察孩子的小便情况,多喝水,避免脱水。如有问题可随时与医生留言,并及时就医。	**Chief complaint**: Child with diarrhea for a week, presenting with symptoms of mild indigestion, sometimes green stools, sometimes foamy stools, sometimes containing white sticky substances. Took Smecta for two days, then went to the hospital, took Attapulgite Powder and lactobacillus for two days, but no improvement. Wondering if the child needs to take antibiotics and what to eat. **Present medical history**: The child has been experiencing diarrhea for a week, characterized by abnormal stool color, sometimes green, sometimes foamy, and sometimes with white sticky substances. The child previously took Smecta, Attapulgite Powder, and lactobacillus, but the symptoms did not improve. **Auxiliary examination**: It is recommended to visit the hospital for a stool examination to rule out infectious enteritis and further determine the cause. **Past medical history**: No relevant medical history. **Diagnosis**: Likely diarrhea due to indigestion. **Recommendation**: It is advised to visit the hospital for a stool examination to determine the cause and receive appropriate treatment based on the examination results. While awaiting the examination results, continue taking probiotics and Attapulgite Powder. Avoid the blind use of antibiotics to prevent harm to the digestive tract. Pay attention to the child's urine situation, drink plenty of water to avoid dehydration. If there are any concerns, feel free to leave a message for the doctor and seek medical attention promptly.

Fig. 18 The main issues with the summaries generated by ChatGPT are: (1). The "Chief Complaint" part is overly lengthy. (2). In the "Auxiliary examination" part, there are suggestions for examinations that did not actually occur. However, despite these issues, they do not affect the understanding of the generated summaries by both the medical professionals and patients

Limitations

Although the dataset used in this article is a publicly available dataset designed for medical natural language processing competitions, which avoids the legal and ethical issues associated with using patient data, protecting sensitive patient data remains a critical area worthy of research and attention. Due to the focus and space limitations of this study, only brief discussions are provided here. We look forward to conducting more detailed research in the future.

It is noteworthy that 87.8% of survey respondents expressed concerns that chatbots could be utilized for data collection or user manipulation [52]. While ChatGPT diligently focuses on ensuring safe conversations and effectively guards against direct prompts used in data extraction attacks during training, there remains a potential vulnerability known as "jailbreaking" that can circumvent its ethical safeguards. As an illustration, ChatGPT may occasionally disclose private details while operating in its "Developer Mode" under a jailbreaking prompt [53]. As the landscape of AI evolves, traditional approaches to information security become outdated. A rule-based strategy is no longer sufficient in the face of generative AI tools [54]. Timo et al. propose that establishing flexible regulatory mechanisms and legal frameworks is crucial, and when regulating the technology and applications of LLMs, it is essential to consider the rapid development of technology and the constantly changing legal environment. Furthermore, cybersecurity vulnerabilities in LLMs can lead to data breaches and malicious attacks, necessitating the establishment of minimum security standards and the provision of appropriate training for healthcare professionals [55].

In conclusion, we believe that relying solely on LLMs providers to protect patient privacy data is insufficient. At the very least, the following key aspects should be considered:

- The continuous improvement of regulatory mechanisms and legal standards permeates the entire process of model creation, deployment, and version updates. Considering the significant costs involved in LLMs training, LLMs providers need to enhance the adaptability of models, especially those that have completed training, in terms of technological innovations and changes in the legal environment.
- Both LLMs providers and data providers must adhere to relevant data security usage standards before inputting medical data into model training, including the use of authorization and authentication tools designed to prevent sensitive information leakage, as well as implementing filtering or encryption measures for medical sensitive data.
- Healthcare institutions need to strictly regulate the use of data and provide rigorous training for healthcare professionals to ensure compliance with relevant laws, behavioral norms, and security standards when using medical data on LLMs.

Abbreviations

GPT	Generative pre-trained transformer;
NLP	Natural language processing;
BERTSUM	Bert-based summarization;
BART	Bidirectional auto-regressive Transformers;
LLMs	Large language models;
ICL	inter-context learning;
BERT	bidirectional encoder representation from Transformers;
UMLS	Unified medical language system;
RLHF	Reinforcement learning from human feedback;
PLMs	Pre-trained language models;
FFN	feedforward network;
PPO	Proximal policy optimization;
CBLUE	Chinese biomedical language understanding evaluation

Acknowledgements
The authors would like to express their gratitude to the School of Data Science at Fudan University for their contribution in creating the Intelligent Medical Consultation System dataset (IMCS-V2), as well as to Alibaba Cloud for providing data storage, download, and task support.

Authors' contributions
YL and SJ designed the research. YL, JW conducted the literature review. YL and JW implemented all methods and performed their experiments. All authors discussed the results. YL, SJ and JW authors wrote the paper. All authors read and approved the final manuscript.

Funding
This work was partially supported by the key project of the National Natural Science Foundation under Grant No. 62137001.

Availability of data and materials
The datasets supporting the conclusions of this study can be found at IMCS-V2-MRG.zip. Prompt Settings related to medical report generated code is placed on https://github.com/gameliu007/Prompt-for-Medical-Report. Additionally, the other relevant source code mentioned in this paper can be referred to in the respective papers of the models.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Computer Science, Sichuan University, No. 24, South Section 1, 1st Ring Road, Chendu 610065, Sichuan, China.

Received: 2 October 2023 Accepted: 11 March 2024
Published online: 14 March 2024

References

1. Jo HS, Park K, Jung SM. A scoping review of consumer needs for cancer information. Patient Educ Couns. 2019;102(7):1237–50.
2. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep. 2019;134(6):617–25.
3. Jain R, Jangra A, Saha S, Jatowt A. A survey on medical document summarization. 2022. arXiv preprint arXiv:2212.01669
4. Navarro DF, Dras M, Berkovsky S. Few-shot fine-tuning SOTA summarization models for medical dialogues. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. 2022. p. 254–266. https://aclanthology.org/2022.naacl-srw.32/.
5. Hollander JE, Carr BG. Virtually perfect? Telemedicine for COVID-19. N Engl J Med. 2020;382(18):1679–81.
6. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc. 2020;27(7):1132–5.
7. Liu Y. Fine-tune BERT for extractive summarization. 2019. arXiv preprint arXiv:1903.10318.
8. Song Y, Tian Y, Wang N, Xia F. Summarizing medical conversations via identifying important utterances. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020. p. 717–29. https://aclanthology.org/2020.coling-main.63/.
9. Krishna K, Khosla S, Bigham JP, Lipton ZC. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. 2020. arXiv preprint arXiv:2005.01795.
10. Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, et al. fairseq: a fast, extensible toolkit for sequence modeling. 2019. arXiv preprint arXiv:1904.01038.
11. Zhang L, Negrinho R, Ghosh A, Jagannathan V, Hassanzadeh HR, Schaaf T, et al. Leveraging pretrained models for automatic summarization of doctor-patient conversations. 2021. arXiv preprint arXiv:2109.12174.
12. Michalopoulos G, Williams K, Singh G, Lin T. MedicalSum: A Guided Clinical Abstractive Summarization Model for Generating Medical Reports from Patient-Doctor Conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2022. 2022. p. 4741–4749.
13. Joshi A, Katariya N, Amatriain X, Kannan A. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. 2020. arXiv preprint arXiv:2009.08666.
14. Mrini K, Dernoncourt F, Chang W, Farcas E, Nakashole N. Joint summarization-entailment optimization for consumer health question understanding. In: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations. 2021. p. 58–65. https://aclanthology.org/2021.nlpmc-1.8/.

15. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. arXiv preprint arXiv:1810.04805.
16. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. 2019. arXiv preprint arXiv:1910.13461.
17. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
18. Ortega-Martín M, García-Sierra Ó, Ardoiz A, Álvarez J, Armenteros JC, Alonso A. Linguistic ambiguity analysis in ChatGPT. 2023. arXiv preprint arXiv:2302.06426.
19. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–44.
20. Giorgi J, Toma A, Xie R, Chen S, An KR, Zheng GX, et al. Clinical Note Generation from Doctor-Patient Conversations using Large Language Models: Insights from MEDIQA-Chat. 2023. arXiv preprint arXiv:2305.02220.
21. Tang X, Tran A, Tan J, Gerstein M. GersteinLab at MEDIQA-Chat 2023: Clinical Note Summarization from Doctor-Patient Conversations through Fine-tuning and In-context Learning. 2023. arXiv preprint arXiv:2305.05001.
22. Ma C, Wu Z, Wang J, Xu S, Wei Y, Liu Z, et al. ImpressionGPT: an iterative optimizing framework for radiology report summarization with chatGPT. 2023. arXiv preprint arXiv:2304.08448.
23. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5(3):e107–8.
24. Dong Q, Li L, Dai D, Zheng C, Wu Z, Chang B, et al. A survey for in-context learning. 2022. arXiv preprint arXiv:2301.00234.
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
26. Graves A. Generating sequences with recurrent neural networks. 2013. arXiv preprint arXiv:1308.0850.
27. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. 2019. arXiv preprint arXiv:1907.11692.
28. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1):5485–551.
29. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019. arXiv preprint arXiv:1910.01108.
30. Sun Z, Yu H, Song X, Liu R, Yang Y, Zhou D. Mobilebert: a compact task-agnostic bert for resource-limited devices. 2020. arXiv preprint arXiv:2004.02984.
31. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced language representation with informative entities. 2019. arXiv preprint arXiv:1905.07129.
32. Moro G, Ragazzi L, Valgimigli L, Freddi D. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. p. 180–9. https://cris.unibo.it/handle/11585/900380.
33. Grail Q, Perez J, Gaussier E. Globalizing BERT-based transformer architectures for long document summarization. In: Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume. 2021. p. 1792–810. https://aclanthology.org/2021.eacl-main.154/.
34. Kieuvongngam V, Tan B, Niu Y. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. 2020. arXiv preprint arXiv:2006.01997.
35. Kanwal N, Rizzo G. Attention-based clinical note summarization. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. 2022. p. 813–20. https://dl.acm.org/doi/abs/10.1145/3477314.3507256.
36. DeYoung J, Beltagy I, van Zuylen M, Kuehl B, Wang LL. Ms2: Multi-document summarization of medical studies. 2021. arXiv preprint arXiv:2104.06486.
37. Gupta S, Sharaff A, Nagwani NK. Biomedical text summarization: a graph-based ranking approach. In: Applied Information Processing Systems: Proceedings of ICCET 2021. Springer; 2022. p. 147–156.
38. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017. arXiv preprint arXiv:1707.06347.
39. Hassani H, Silva ES. The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. Big Data Cogn Comput. 2023;7(2):62.
40. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. J Assoc Inf Sci Technol. 2023;74(5):570–81.
41. Abdullah M, Madain A, Jararweh Y. ChatGPT: Fundamentals, applications and social impacts. In: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE; 2022. p. 1–8.
42. Baidoo-Anu D, Owusu Ansah L. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. J AI. 2023;7(1):52–62. https://dergipark.org.tr/en/pub/jai/issue/77844/1337500.
43. Jiao W, Wang W, Huang Jt, Wang X, Tu Z. Is ChatGPT a good translator? A preliminary study. 2023. arXiv preprint arXiv:2301.08745.
44. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33.
45. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.
46. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. Clin Transl Med. 2023;13(3). https://doi.org/10.1002/ctm2.1216.
47. Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. Am J Roentgenol. 2023. https://doi.org/10.2214/AJR.23.29198.
48. Shaib C, Li ML, Joseph S, Marshall IJ, Li JJ, Wallace BC. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). 2023. arXiv preprint arXiv:2305.06299.
49. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-Related research and perspective towards the future of large language models. 2023. arXiv preprint arXiv:2304.01852.
50. Lin CY. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. Barcelona: Association for Computational Linguistics; 2004. p. 74–81. https://aclanthology.org/W04-1013.
51. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: evaluating text generation with bert. 2019. arXiv preprint arXiv:1904.09675.
52. Sebastian G. Do ChatGPT and other AI chatbots pose a cybersecurity risk?: An exploratory study. Int J Secur Priv Pervasive Comput. 2023;15(1):1–11.
53. Li H, Guo D, Fan W, Xu M, Song Y. Multi-step jailbreaking privacy attacks on chatgpt. 2023. arXiv preprint arXiv:2304.05197.
54. Renaud K, Warkentin M, Westerman G. From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI. MIT Sloan Management Review; 2023.
55. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA. 2023. https://doi.org/10.1001/jama.2023.9651.

Publisher's Note