



# Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers

Jiajia Yuan,<sup>1,8</sup> Peng Bao,<sup>2,8</sup> Zifan Chen,<sup>2,8</sup> Mingze Yuan,<sup>2,8</sup> Jie Zhao,<sup>3,7</sup> Jiahua Pan,<sup>7</sup> Yi Xie,<sup>1</sup> Yanshuo Cao,<sup>1</sup> Yakun Wang,<sup>1</sup> Zhenghang Wang,<sup>1</sup> Zhihao Lu,<sup>1</sup> Xiaotian Zhang,<sup>1</sup> Jian Li,<sup>1</sup> Lei Ma,<sup>6</sup> Yang Chen,<sup>1,\*</sup> Li Zhang,<sup>2,6,\*</sup> Lin Shen,<sup>1,\*</sup> and Bin Dong<sup>4,5,6,7,\*</sup>

\*Correspondence: yang\_chen@bjcancer.org (Y.C.); zhangli\_pku@pku.edu.cn (L.Z.); shenlin@bjmu.edu.cn (L.S.); dongbin@math.pku.edu.cn (B.D.)

Received: July 24, 2023; Accepted: August 8, 2023; Published Online: August 14, 2023; <https://doi.org/10.59717/j.xinn-med.2023.100019>

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## GRAPHICAL ABSTRACT



## PUBLIC SUMMARY

- Prompt engineering affects large language models' performance in GI oncology.
- Prompts with templates and in-context learning enhance large language models' output.
- Multi-round interaction helps large language models to reach the best performance.
- Such performance meets the need of senior GI oncologists for effective AI agents.



# Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers

Jiajia Yuan,<sup>1,8</sup> Peng Bao,<sup>2,8</sup> Zifan Chen,<sup>2,8</sup> Mingze Yuan,<sup>2,8</sup> Jie Zhao,<sup>3,7</sup> Jiahua Pan,<sup>7</sup> Yi Xie,<sup>1</sup> Yanshuo Cao,<sup>1</sup> Yakun Wang,<sup>1</sup> Zhenghang Wang,<sup>1</sup> Zhihao Lu,<sup>1</sup> Xiaotian Zhang,<sup>1</sup> Jian Li,<sup>1</sup> Lei Ma,<sup>6</sup> Yang Chen,<sup>1,\*</sup> Li Zhang,<sup>2,6,\*</sup> Lin Shen,<sup>1,\*</sup> and Bin Dong<sup>4,5,6,7,\*</sup>

<sup>1</sup>Department of Gastrointestinal Oncology, Key Laboratory of Carcinogenesis and Translational Research, Peking University Cancer Hospital and Institute, Beijing 100142, China

<sup>2</sup>Center for Data Science, Peking University, Beijing 100871, China

<sup>3</sup>National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871, China

<sup>4</sup>Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China

<sup>5</sup>Center for Machine Learning Research, Peking University, Beijing 100871, China

<sup>6</sup>National Biomedical Imaging Center, Peking University, Beijing 100871, China

<sup>7</sup>Peking University Changsha Institute for Computing and Digital Economy, Changsha 410205, China

<sup>8</sup>These authors contributed equally

\*Correspondence: yang\_chen@bjcancer.org (Y.C.); zhangli\_pku@pku.edu.cn (L.Z.); shenlin@bjmu.edu.cn (L.S.); dongbin@math.pku.edu.cn (B.D.)

Received: July 24, 2023; Accepted: August 8, 2023; Published Online: August 14, 2023; <https://doi.org/10.59717/j.xinn-med.2023.100019>

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Citation: Yuan J, Bao P, Chen Z, et al., (2023). Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers. *The Innovation Medicine* 1(2), 100019.

Large Language Models' (LLMs) performance in healthcare can be significantly impacted by prompt engineering. However, the area of study remains relatively uncharted in gastrointestinal oncology until now. Our research delves into this unexplored territory, investigating the efficacy of varied prompting strategies, including simple prompts, templated prompts, in-context learning (ICL), and multi-round iterative questioning, for optimizing the performance of LLMs within a medical setting. We develop a comprehensive evaluation system to assess the performance of LLMs across multiple dimensions. This robust evaluation system ensures a thorough assessment of the LLMs' capabilities in the field of medicine. Our findings suggest a positive relationship between the comprehensiveness of the prompts and the LLMs' performance. Notably, the multi-round strategy, which is characterized by iterative question-and-answer rounds, consistently yields the best results. ICL, a strategy that capitalizes on interrelated contextual learning, also displays significant promise, surpassing the outcomes achieved with simpler prompts. The research underscores the potential of advanced prompt engineering and iterative learning approaches for boosting the applicability of LLMs in healthcare. We recommend that additional research be conducted to refine these strategies and investigate their potential integration, to truly harness the full potential of LLMs in medical applications.

## INTRODUCTION

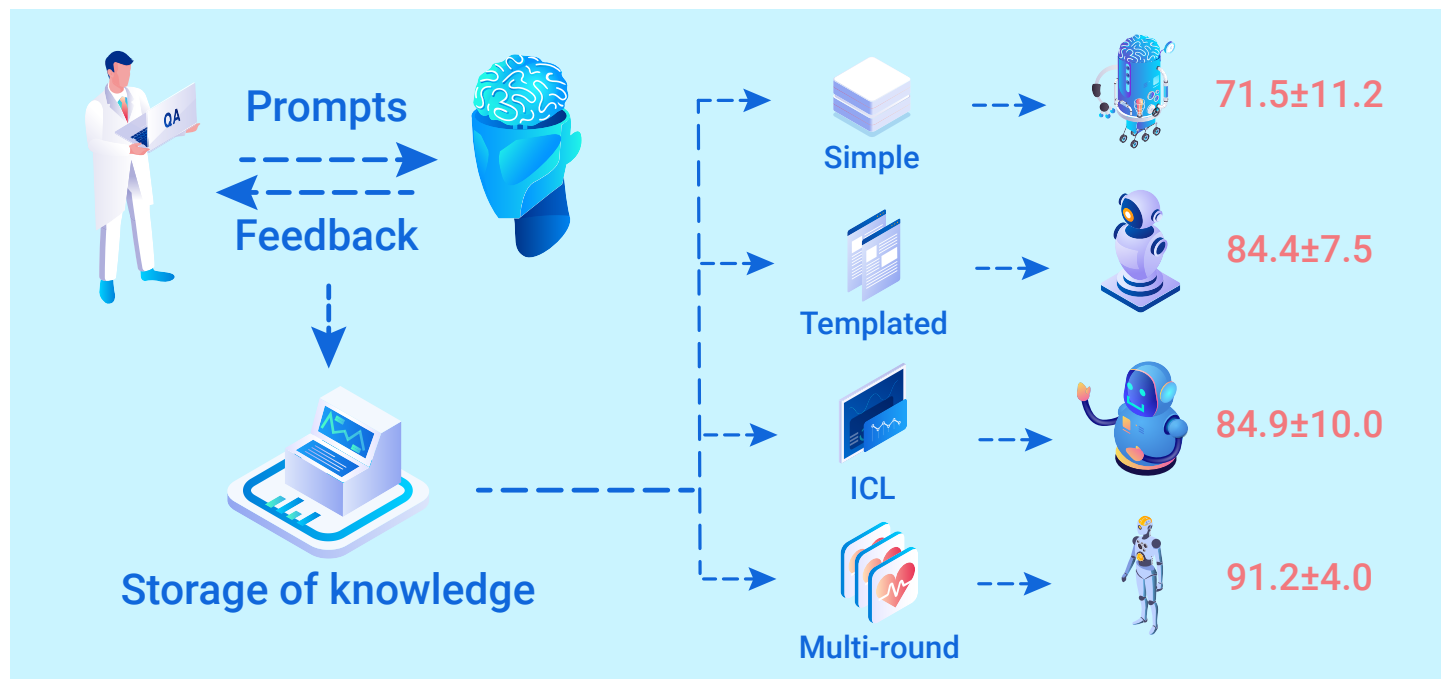
Large Language Models (LLMs), exemplified by cutting-edge architectures like GPT-4,<sup>1</sup> have demonstrated considerable potential in transforming healthcare delivery<sup>2-4</sup> and competency in medical examinations.<sup>5</sup> This influence is manifested across various healthcare sectors, including online patient interaction,<sup>6</sup> preventive oncology,<sup>7-9</sup> neuropsychiatry,<sup>10</sup> dermatology,<sup>11</sup> and aesthetic surgery consultation,<sup>12,13</sup> underscoring their remarkable versatility. However, the application of LLMs such as GPT-4 in digestive system cancer treatment remains an underexplored area. The complexities inherent to this field, from patient consultation, and diagnosis to treatment planning and follow-up care, pose formidable challenges for LLMs. Additionally, the existing body of research<sup>2,5,7</sup> primarily evaluates LLMs' responses to common medical inquiries via rudimentary prompting, which may not fully leverage their potential in medical settings. This highlights the need for a more comprehensive assessment of GPT-4's capability to provide personalized cancer treatment recommendations via sophisticated prompts.

To harness the full potential of LLMs, it is crucial to employ effective prompt engineering.<sup>14-19</sup> Prompt engineering, a process of creating, testing, and optimizing input prompts, serves as a crucial tool in controlling and enhancing interactions with LLMs. Various techniques such as in-context learning,<sup>15</sup> retrieval-augmented generation,<sup>16</sup> chain-of-thought,<sup>17</sup> and least-to-most prompting<sup>18</sup> have been shown to significantly improve the performance of LLMs in tasks demanding logical thinking and reasoning. In-context learn-

ing offers models a few demonstrations before attempting a task, while retrieval-augmented generation enhances this process by retrieving relevant examples from a given database. Chain-of-thought prompting improves LLMs' reasoning ability by directing them to generate a series of intermediate steps toward a solution, and least-to-most prompting dissects complex problems into simpler sub-problems to be solved sequentially. Intuitively, these techniques could effectively boost LLMs' performance in complex medical tasks, including cancer treatment recommendations.

In this study, we aimed to unleash GPT-4's potential to provide personalized digestive system cancer treatment plans through prompt engineering. Inspired by the thinking, reasoning, and action processes of digestive oncologists, we initially conceived the iterative procedure of prompt engineering as a method of amassing information regarding gastrointestinal tumors within a distinct storage of knowledge and in turn, educating the LLM. However, these knowledge repositories, when embedded in rudimentary prompts, are often devoid of substantial content, thus limiting their potential to effectively guide LLMs. Consequently, we established an empirically effective multi-step prompt template consisting of: (i) declaring the role, this process involves assigning a particular role to GPT-4 that emulates a real-world professional or function; (ii) stating the main task, this step essentially provides GPT-4 with a clear directive of what it is required to accomplish; (iii) declaring the workflow, which we view as a generalized chain of thought that allows GPT-4 to approach problem-solving or deliver answers in an organized, step-by-step manner; (iv) specifying constraints, it involves defining the boundaries within which GPT-4 should operate. Then, we iteratively refined this template to align GPT-4's responses with physicians' requirements and added elements to generate comforting responses for patients. An experienced oncology specialist subsequently interacted with GPT-4 over multiple rounds to further guide and optimize the recommended treatment plans. Furthermore, motivated by the exemplar-based teaching approach in medicine, we also assessed the impact of in-context learning by providing GPT-4 with examples of ideal treatment suggestions through document retrieval. We evaluated the performance of diverse prompt engineering strategies on 43 case reports, encompassing a wide range of digestive system cancer types, utilizing a clinically standardized evaluation metric.

In summary, we are the first to conduct a comprehensive assessment of prompt engineering on GPT-4's ability to provide personalized digestive system cancer treatment recommendations, as per our comprehensive search in the existing literature. We developed a sophisticated prompt template to generate personalized cancer treatment plans that emphasize patient comfort, which significantly outperforms rudimentary prompts and offers valuable insights for prompt design in the medical domain. We evaluated various prompt engineering strategies, including rudimentary prompts, templated prompts, in-context learning, and multi-round interaction, using a clinically standardized metric. Our results highlight the promise of prompt engineering for medical applications of LLMs.



**Figure 1.** An illustration showcasing the effects of various prompting strategies on Language Learning Models' (LLMs) performance, mediated by a 'storage of knowledge'. Simple prompts leave this storage empty, offering no enhancement for GI tumor decision-making. Conversely, templated prompts and ICL populate the storage with role assumptions and case examples, respectively, helping to standardize LLMs' output, thus improving performance. The multi-round interaction strategy fills the storage with the complete physician-LLM dialogue, potentially allowing more accurate comprehension and utilization of decision-assisting information.

## MATERIALS AND METHODS

### Materials

In this study, we propose an innovative methodology to augment the learning capability of LLMs by incorporating multifaceted prompt design and dynamic training approaches. As shown in Figure 1, diverse prompt designs can be perceived as varying modifications to the storage of knowledge, encompassing manual alterations meticulously orchestrated based on GI tumor expertise, automatic modifications that explore the hospital's pre-existing data for analogous cases as pedagogical instances for the LLM, and dynamic modifications consistently interrogated and addressed during the deployment of the consultation process. Consequently, the design of the prompts was executed as follows: Initially, the models are subjected to a more sophisticated introduction prompt, intricately crafted with complex semantic and structural nuances, thereby priming the LLMs to comprehend and respond to intricate queries. Furthermore, an advanced method of in-context learning is introduced, encouraging the models to extract knowledge and patterns from various contexts rather than individual sentences, fostering a more comprehensive understanding of the text. To accommodate evolving data patterns, we also incorporate online learning techniques, enabling the LLMs to continually learn and adapt from real-time, dynamic data. Lastly, we implement an iterative feedback loop through multi-round question-and-answer sessions, reinforcing the model's ability to comprehend, retain, and apply information over successive interactions. This combination of sophisticated prompt architecture, in-context learning, online learning, and iterative interactions aims to substantially enhance the LLM's predictive and interpretative capabilities, pushing the frontiers of AI language understanding. We used publicly available medical licensing examination cases, oncology residency and attending physician exam cases as text source.

### Templated prompts

Past studies have shown that a good use of different prompt engineering,<sup>17,20,21</sup> as well as properly designed prompt templates<sup>22</sup> can significantly improve the problem-solving ability of large language models, and this phenomenon was similarly observed in our study. As shown in Figure S1, we developed our prompt template by adopting a four-pronged approach as follows:

**Declaring the role.** Assigning a 'role' or 'identity' to large language models is one of the commonly used techniques for interacting with these models. Previous research<sup>22</sup> supports that this method can effectively guide what type of output the models generate and what details they prioritize. In our study, we assigned the role of a digestive oncology specialist to GPT-4, emphasizing its range of skills that included clinical diagnosis, treatment, and communication techniques. We found this strategy successfully influenced GPT-4's behaviors, responses, and interaction styles to align with the expectations of the role.

**Stating the main task.** This approach essentially provides GPT-4 with a clear directive on what it is expected to accomplish. In our study, the primary task of our model is to deliver detailed and accurate advice to patients with digestive system cancers. This involves defining the central task that GPT-4 needs to perform. Given the context of our research, our model, acting as a digestive oncology specialist, is tasked with generating personalized treatment plans for digestive system cancer patients. By articulating the main task, we direct GPT-4's focus, streamline its reasoning process, and enhance its ability to produce task-specific, relevant, and actionable outputs. In addition, to enable GPT-4 to produce complex and contextually accurate responses, we've included a wide range of scenarios and contexts, from simple situations to the complexity of academic discourse in hospitals. We also encourage GPT-4 to link different pieces of information together. This approach aids GPT-4 in moving beyond simple pattern recognition, facilitating a deeper understanding when executing tasks.

**Declaring the workflow.** We have defined a comprehensive workflow in the prompt templates, which includes case analysis, clinical examination, scheduling examination, diagnosis and treatment, execution and adjustment of treatment, and follow-ups. This is also the general workflow of a professional digestive oncology specialist. We believe this represents a generalized chain of thought and many studies<sup>17,20,21,23</sup> have already demonstrated that this approach can stimulate LLM's reasoning ability. We find this strategy ensures that GPT-4's output is more consistent and logical, using a planned, step-by-step approach to accomplish tasks, which is very similar to the process a human expert uses to solve problems. By structuring GPT-4's thinking in this way, we can effectively manage its output, improve overall consistency, and reduce the likelihood of generating irrelevant or erroneous information.

**Specifying constraints.** In this process, we've incorporated certain constraints into the prompt templates. We require GPT-4 not to make responses when uncertain or additional information is needed, but rather, it must first gather sufficient information. In addition, we require GPT-4 to provide detailed and correct guidance for a specific case, as GPT-4 tends to give general and non-specific answers that may not be wrong but lack specificity. This approach ensures that GPT-4 avoids generating responses that are undesirable or beyond its scope, thereby enhancing its effectiveness and minimizing potential deviations. We also advised GPT-4 to build a trusting doctor-patient relationship in a warm, humorous manner rather than in a cold and impersonal way when answering.

### In-context learning

In this study, we introduce an automated in-context learning (ICL) approach to refine GPT-4's capabilities, focusing on the integration of doctors' habits and cognition. This method assimilates insights drawn from analogous past cases and is comprised of three main components: firstly, transposing past patient conditions into a designated embedding space; secondly, gauging the similarity between the current condition and these archived cases to identify its k-nearest counterparts; and finally, building in-context learning prompts based on these identified cases. We provide a detailed exposition of these three components in the following:

**Encoding patient conditions using pre-trained chinese BERT model.** A pre-trained Chinese BERT model in Hugging Face (<https://huggingface.co/hfl/chinese-bert-wwm-ext>), specifically the "hfl/chinese-bert-wwm-ext", is utilized to translate patient conditions into a high-dimensional embedding space (768 dimensions in this study), capturing the context of the condition effectively. The BERT tokenizer is used to convert condition text into input vectors, which are then fed into the BERT model. Operating in a no-gradient update setting, the "pooler\_output" from the model serves as the sentence embedding for each patient condition.

**Calculation of cosine similarity and identification of k-nearest neighbors.** Once the embeddings for all patient conditions have been computed, we calculate the cosine similarity between them to derive a similarity score. This metric provides a measure of the contextual similarity between different patient conditions. Based on these similarity scores, we identify up to k-nearest neighbors for each patient condition (with k being up to four depending on the token limitation of GPT-4).

**Generation of in-context learning prompts.** For each patient condition, we generate an enriched prompt that includes the top-k similar past cases and the corresponding doctor's suggestions. To ensure consistency and readability in these prompts, a pre-defined template is used: "As an experienced clinician, your responsibilities include understanding and analyzing patient information and chief complaints, [...]. Now, let's look at these examples: [...]. After analyzing these examples, here is a new patient: [...]. Please give specific treatment plan suggestions based on the above examples and relevant literature. (see [Figures 4 & S6](#) for details).

### Metrics

We have developed a unique set of metrics, drawing from those typically used for evaluating clinicians' examinations, to quantitatively assess the results generated by various methods. These metrics encompass six key aspects:

**1. Understanding medical history (0-20):** This metric assesses how accurately and comprehensively an LLM captures and interprets a patient's medical history. This includes consideration of the patient's previous diagnoses, surgeries, hospitalizations, allergies, family history, lifestyle, and other relevant information.

**2. Diagnosis and differential diagnosis (0-20):** This metric assesses the ability of the LLM to accurately diagnose the patient's condition based on the medical history. It includes both the primary diagnosis and any differential diagnoses.

**3. Further examination and reason (0-10):** This metric evaluates the appropriateness of any additional examinations suggested by the LLM. It measures not only whether the recommended examinations are suitable, but also if they are justified based on the patient's condition and symptoms. The LLM

should also provide a clear rationale for why these additional examinations are needed.

**4. Principles and plans of treatment (0-20):** This metric evaluates the LLM's ability to propose a suitable treatment plan. The plan should be personalized for the patient, taking into account factors like age, overall health, potential side effects, and patient preferences.

**5. Breadth and depth of results (0-20):** This metric measures how comprehensively the LLM covers the scope of medical knowledge in its results (breadth), as well as how much detail it provides (depth). Breadth refers to the range of different topics or areas covered in the results, while depth refers to the level of detail or complexity within those topics.

**6. Thinking and expressing ability (0-10):** This is a measure of how effectively the LLM reasons and communicates its findings. Thinking refers to the LLM's ability to logically process and interpret data, make connections, draw conclusions, and anticipate potential outcomes. The expressing ability should not only be clear and accurate but also demonstrate empathy in line with a real clinician's interaction. This includes sensitivity to the patient's emotional state, using comforting and supportive language, and showing understanding and respect for the patient's experiences and concerns. By effectively incorporating empathy, the LLM can build trust, encourage open communication, and provide emotional support in addition to addressing physical health issues.

To gain a clearer understanding of performance based on the total scores, we have defined the following expertise levels:

**1. Level A (90-100 points):** Top-level expertise, capable of independently managing complex and rare cases, demonstrating exceptional skills and professional knowledge.

**2. Level B (80-89 points):** Experienced level, capable of handling most cases, but requires guidance for complex or rare cases.

**3. Level C (70-79 points):** Mid-level competence, capable of independently addressing common cases, requires guidance for complex ones.

**4. Level D (60-69 points):** Junior level, capable of handling some common cases, but requires close guidance for complex cases.

**5. Level E (below 60 points):** Initial training level, needs guidance from experienced clinicians in all aspects.

## RESULTS

### Templated evaluation

[Figures 2 & S2](#) provide a comparison between our designed prompting template ([Figure 2B](#)) and the standard, direct prompting ([Figure 2A](#)) utilized by GPT-4. The findings underscore that the designed template for role assumption ([Figure S1](#)) can improve GPT-4 to make more complex decisions based on the patient's individual circumstances. In the provided example, our designed prompting can prioritize the control of disease progression, symptom relief, enhancement of life quality, and survival extension, instead of merely pursuing a cure unconditionally. Moreover, the template manifests an exceptional ability to interweave quality-of-life considerations within the treatment strategies and provides comprehensive guidance ([Figure S3](#)). It also underscores the significance of continuous patient assessment and the pursuit of innovative, custom treatments ([Figure S4](#)). As opposed to direct prompting, our designed prompting template possesses the ability to mimic the intricate treatment ideation process, enhancing GPT-4's efficacy as a therapeutic advisory tool when acting as a senior oncologist.

### Multi-round evaluation

[Figures 3 & S5](#) illustrate an interaction with GPT-4 for cancer treatment advice. Initially, GPT-4 prematurely diagnosed the patient with late-stage cancer and proposed a treatment plan. However, this was inappropriate, given the necessity for a more accurate staging diagnosis for this patient. As highlighted in [Figure 3](#), the clinician directed GPT-4 to offer a detailed staging diagnosis, subsequently pointing out its error. Following multiple question-and-answer interactions with the clinician, GPT-4 acknowledged its mistake and adjusted its response. It began by determining the cancer's stage, before suggesting a specific treatment plan. This revised response is not only more suitable for the patient but also provides her with hope. This multi-round interaction demonstrates the learning capability of large language models like GPT-4, highlighting their ability to quickly integrate human logical reasoning within the context of intricate medical scenarios.

## A Direct prompting



Doctor

Male, 68 years old. The patient was found to have a gastric mass in an external hospital. Gastroscopy showed that at 41-47cm from the incisors, there is an ulcerative tumor at the esophagogastric junction. Pathology: Esophagogastric junction. Biopsy: moderately differentiated adenocarcinoma. Lauren classification: intestinal type. Immunohistochemical results: HER2(0) [HER2 positive control (3+), HER2 negative control (0)]. CT: Thickening of the gastric wall, considered as Borrmann type 3 gastric cancer(ct4aN3a) ; Multiple filling defects in the splenic vein and portal vein, suggesting tumor thrombus. The tumor markers are normal. Considering the patient's multiple portal vein tumor thrombi, it is not suitable for clinical trial enrollment.

This is the information of a patient. Please search for relevant literature and provide specific treatment plan recommendations, and explain the reasons.

- This prompt is summarized into "{Sex}, {Age}, {Disease description}, {Question}." and will be referred to in other content.

Based on the information provided, immunohistochemistry shows HER2 positive. Due to the presence of multiple tumor thrombi in the portal vein, the patient is not suitable for clinical trial enrollment.

The treatment plan is as follows:

1. Chemotherapy: Considering the patient's condition, neoadjuvant chemotherapy can be administered to reduce tumor size, alleviate symptoms, and increase the possibility of surgery. Common chemotherapy regimens include the FLOT regimen and the EOX regimen.
2. Surgery: If the tumor shrinks and the condition stabilizes after chemotherapy, the possibility of surgical removal can be evaluated. Depending on the patient's specific situation, a gastric cancer radical operation (D2 or D1+ lymph node dissection) or palliative surgery can be chosen.
3. Targeted therapy: For HER2-positive patients, HER2-targeted therapy drugs such as trastuzumab can be added during chemotherapy.
4. Supportive therapy: Nutritional support and symptomatic treatment.



GPT-4

## B Make role assumptions before prompting



Doctor

{Role declaration}  
 {Definition of a Gastrointestinal Oncologist}  
 {Skills of a Gastrointestinal Oncologist}  
 {Work Process of a Gastrointestinal Oncologist}  
 {Specific Requirements and Rules for the Response}

- {Sex}, {Age}, {Disease description}.

Based on the information you provided, the patient has been diagnosed with moderately differentiated adenocarcinoma of the esophagogastric junction, with a pathological type of intestinal, HER2-negative. CT scans show the existence of tumor thrombus, which may suggest the disease is in the late stage. However, please note that a more detailed clinical data and expert evaluation are needed for a definite diagnosis.

For patients with advanced gastric cancer of this type, systemic chemotherapy is often the first choice, such as the FOLFOX regimen (fluorouracil + cisplatin + oxaliplatin) or the SOX regimen (S-1 + oxaliplatin). Moreover, considering the HER2-negative pathological result, targeted drugs like trastuzumab may not be applicable here. Additionally, if the disease is well-controlled, palliative surgery can be considered to alleviate symptoms.

Regarding the tumor thrombus, we need to evaluate in detail whether there is a possibility for surgery or interventional treatment, or if anticoagulant therapy is needed. If the tumor thrombus causes severe clinical symptoms, surgery or interventional treatment may be considered, but this requires a comprehensive assessment of the patient's overall condition and other complications.

At the same time, it is also recommended to conduct relevant genetic testing, such as MLH1, MSH2, MSH6, PMS2, etc., because a family history of cancer may be related to hereditary cancer syndromes.

This is a possible diagnosis and treatment plan based on the information I currently have. The actual treatment plan needs to be made based on the patient's specific situation and a comprehensive assessment by the medical team. Please discuss these suggestions with your medical team to develop the most suitable treatment plan for the patient.



GPT-4

**Figure 2. Evaluation of prompting template design** (A) Direct prompting. GPT-4 is directly provided with a patient's condition and asked to suggest a treatment plan. (B) The proposed templated prompting (make role assumptions before prompting). GPT-4 is initially primed with a specifically designed role-embedded prompt, ensuring the AI understands the task requirements and responsibilities. Subsequently, the patient's condition is provided, and GPT-4 is asked to suggest a treatment plan. Note that this figure is a reduced version of [Figure S2](#).

### ICL evaluation

As demonstrated in [Tables 1-4](#), the performance of In-Context Learning (ICL) exceeded that of rudimentary prompting by a substantial margin across

various types of digestive system cancer treatments, with a notable difference of 13.4 points in overall performance. [Figures 4 & S6](#) provide an illustrative comparison between in-context learning and rudimentary prompting.



Doctor

**{Role assumption first}**

Female, 76 years old. Poor appetite. Gastroscopy revealed a raised lesion with a concave surface at the cardia, extending to the gastric fundus (Siewert II type, adenocarcinoma), and coarse gastric mucosa consistent with Borrmann IV type gastric cancer presentation (adenocarcinoma). Pathology report: a small amount of poorly differentiated cancer (gastric body), poorly differentiated adenocarcinoma; Cmet(1+), EGFR(0), HER2(0), pMMR, pan-TRK(-), PD-L1 (CPS<1), EBER(-); (Cardia) moderately differentiated adenocarcinoma, intestinal type; Cmet(2+), EGFR(1+), HER2(0), pMMR, pan-TRK(-), PD-L1 (CPS=10), EBER(-). Normal lab results, chest-abdominal-pelvic enhanced CT showed diffuse thickening of the entire stomach, lymph node metastasis between the liver and stomach. Currently, the patient's food intake has decreased, bowel movements are not abnormal, and she has lost about 5 kg in the past 2 months.

Based on the above information, the treatment plan should focus on alleviating symptoms and improving quality of life. Specific suggestions are as follows:

1. Chemotherapy: Common chemotherapy regimens include FOLFOX, XELOX, SOX, etc.
2. Targeted Drugs: C-MET and EGFR are slightly positive, so targeted drugs against these two markers, such as crizotinib (a C-MET inhibitor) and erlotinib (an EGFR inhibitor) can be considered.
3. Immunotherapy: Although the expression level of PD-L1 is low, immunotherapy checkpoint inhibitors like Keytruda (pembrolizumab) can still be considered.
4. Palliative Care: This includes pain management, nutritional support, and psychosocial support.
5. Regular Follow-up: Regular gastroscopy review and chest, abdominal and pelvic CT scans are necessary.



GPT-4



Doctor

How do you analyze and determine the staging diagnosis and future prognosis for this patient?

The staging of gastric cancer primarily refers to the American Joint Committee on Cancer (AJCC) TNM staging system, which assesses the depth of tumor invasion (T), the status of lymph node metastasis (N), and the condition of distant metastasis (M). According to the information you provided, the patient's gastric cancer has developed lymph node metastasis (LNM), indicating that the N part of the TNM stage is at least N1 (1-2 regional lymph nodes metastasis) or higher. Simultaneously, the gastric cancer is Borrmann IV type, suggesting that the tumor has extensively invaded the stomach wall, and in the TNM stage, the grading of the T part is usually T3 (tumor invades the serosa) or T4 (tumor penetrates the serosa or directly invades adjacent structures). Based on this information, it can be preliminarily judged that the patient is in at least stage III of gastric cancer (including T3N1M0, T4aN1M0, T4bN1M0, etc.) or a higher stage.



GPT-4



Doctor

The patient's lymph node metastasis (LNM) between the liver and stomach could be classified as either N1 or N3, and a detailed interpretation of the lymph node metastasis is needed. Will there be a chance for surgery in the future? This will have a significant impact on future treatment outcomes.

In terms of the resectability of the disease, the patient's lymph node metastasis between the liver and stomach (LNM) may affect surgical decisions. The N stage in the TNM staging is at least N1 (1-2 regional lymph node metastases). In some cases, for locally advanced gastric cancer, even with lymph node metastasis, radical surgery (such as D2 gastrectomy) may be performed. This surgery can remove as many lymph nodes as possible, but the risk of surgery is high, and the rate of postoperative complications and mortality is also relatively high, especially in the elderly population. However, for advanced gastric cancer (such as the patient's Borrmann IV type gastric cancer), especially in cases accompanied by extensive lymph node metastasis or distant metastasis, surgery is usually not the first choice of treatment, because surgery at this time cannot significantly improve prognosis and may instead bring greater risks and complications.



GPT-4



Doctor

Based on the above discussion, please provide a new recommendation for this patient.

I suggest the following:

1. Detailed Staging: First, I suggest more detailed tumor staging. This may include PET-CT scanning to find any other unknown metastatic lesions. Additionally, for lymph node metastasis, we need to further clarify its extent in order to stage more accurately.
2. Peritoneal Metastasis Examination: Considering the possibility of common peritoneal metastasis in gastric cancer, I recommend laparoscopic examination for direct observation of the abdominal cavity and to obtain living tissues for pathological examination, to find any potential hidden peritoneal metastasis.
3. Individualized Treatment Plan: Based on the detailed staging and peritoneal metastasis examination results, we can further discuss the treatment plan. This may include chemotherapy, targeted therapy, and immunotherapy. Especially, immunotherapy might be tried considering her PD-L1(22C3)(CPS=10).
4. Nutrition and Symptom Management: Given her age and current condition (weight loss of about 5kg in the past 2 months, decreased food intake), I recommend asking a dietitian for evaluation and management, and there may also need to be appropriate palliative treatment for symptoms such as decreased appetite.



GPT-4

**Figure 3. A Multi-round Interaction with GPT-4 for Gastric Cancer Treatment Advice** Note that this figure is a reduced version of [Figure S5](#).

## A Direct prompting



Doctor

Male, 59 years old. Visited the clinic due to abdominal pain after eating. Colonoscopy indicated a tumor 30cm from the hepatic flexure. Pathology: Sigmoid colon: Moderately differentiated adenocarcinoma with mucinous adenocarcinoma. PET-CT: Considering colon cancer, multiple metastatic lymph nodes in the abdominal pelvic cavity, peritoneum, and retroperitoneum, multiple liver metastases. Tumor markers: CEA 37.93, CA199 368.3, CA724 103.4, CA242 463.6. Recent weight loss of 8KG.  
This is the information of a patient. Please search for relevant literature and provide specific treatment plan recommendations, and explain the reasons.

● This prompt is summarized into "{Target case}, {Question}." and will be referred to in other content.

- 1. Chemotherapy:** Common chemotherapy drugs include fluorouracil (5-FU), oxaliplatin, and irinotecan. Chemotherapy regimens can use FOLFOX or FOLFIRI.
- 2. Targeted therapy:** Based on the genetic testing results of the patient's tumor, you can consider using drugs that target specific targets for treatment, such as EGFR inhibitors (such as Herceptin, cetuximab or panitumumab), VEGF inhibitors (such as bevacizumab) or other applicable targeted therapy drugs.
- 3. Immunotherapy:** PD-1/PD-L1 inhibitors (like pembrolizumab or nivolumab) may be effective for some patients. The applicability of immunotherapy needs to be determined based on the patient's tumor microsatellite instability (MSI) or PD-L1 expression.
- 4. Supportive care:** During treatment, the patient may need hydration, nutritional support, pain relief, and antibiotic treatment, among other supportive care.
- 5. Follow-up and monitoring:** During treatment, tumor markers should be monitored regularly to assess the effectiveness of the treatment.



GPT-4

## B In-context learning prompting



Doctor

As an experienced clinical physician, your responsibilities include understanding and analyzing patient information and chief complaints, then based on previous cases and the treatment inspection suggestions of real doctors in those cases, give professional treatment inspection suggestions in a similar style. You need to be professional, careful, and thoughtful, and your answer must be based on existing medical knowledge and consider the patient's personal situation and needs. Now, let's look at these examples:

Case 1: {Case description 1}

Doctor's suggestion: {Doctor's suggestion 1}

Case 2: {Case description 2}

Doctor's suggestion: {Doctor's suggestion 2}

After analyzing these examples, here is a new patient: [Target Case in Panel A]. Please give specific treatment plan suggestions based on the above examples and relevant literature.

Based on the information you provided, I suggest the following treatment plan:

1. Complete gene testing, including KRAS/NRAS/BRAF gene testing, as well as immunohistochemistry of MLH1/PMS2/MSH2/MSH6 to clarify the molecular characteristics of the tumor. If it is dMMR/MSI, it is recommended to treat with PD1 monotherapy ± CTLA4 monotherapy.
2. Choose the appropriate chemotherapy plan based on the results of gene testing. If it is a KRAS/NRAS/BRAF wild type, consider the FOLFIRI scheme combined with targeted therapy (such as cetuximab or bevacizumab). If the gene testing results show mutations, select the FOLFIRI scheme or other suitable chemotherapy plans.
3. Strengthen nutritional support in conjunction with the patient's recent weight loss, to improve the patient's nutritional status.
4. Regularly review tumor markers and imaging examinations (such as CT or MRI of the peritoneum, retroperitoneal lymph nodes, and liver) to assess the treatment effect.
5. Based on the treatment effect and the patient's physical condition, assess the possibility of local treatment, such as surgery, ablation, or radiation therapy.
6. If conditions permit, consider participating in clinical research to obtain more treatment opportunities.
7. Pay attention to observe the patient's symptoms and side effects, and conduct symptomatic treatment accordingly.



GPT-4

**Figure 4. Utilizing GPT-4 to Suggest Oncological Treatment Regimens with Distinct Prompting Techniques** (A) Direct prompting and (B) In-Context Learning (ICL) prompting. Note that GPT-4 was not provided with the designed prompt, as our objective was to exclusively evaluate the performance of ICL. Note that this figure is a reduced version of Figure S6.

Panel A portrays a clinical scenario in which a patient is diagnosed with rectal cancer and is accompanied by a naive prompt for generating treatment strategy recommendations. In contrast, Panel B highlights the implementation of in-context learning, where GPT-4 is provided with two comparable patient cases, each accompanied by treatment suggestions endorsed by an experi-

enced oncologist. Through in-context learning, GPT-4 successfully absorbed the treatment approach and reasoning process utilized by physicians, leading to the generation of personalized and targeted treatment plans, as opposed to general guideline recommendations offered through rudimentary prompts. Furthermore, ICL emphasized clinically critical aspects such as clin-

ical research and potential side effects, thus enhancing the overall quality and relevance of the generated treatment recommendations.

### Overall evaluation

Tables 1-4 provided the performance evaluation of different prompt engineering strategies-Simple, Templated, ICL, and Multi (Multiple Rounds)-on several aspects of understanding and knowledge organization across different disease conditions (Overall, Gastric Cancer, Colorectal Cancer, Other GI

cancers). Broadly, the tables show a general trend of increased performance as we move from the Simple strategy to the Multi-strategy, but templated prompts and ICL show similar performance. The mean scores for all aspects improve noticeably as the complexity of the prompts increases as shown in Table S1.

A few specific observations can be highlighted. First, 'Understanding Medical History' consistently receives full marks in the Multi-strategy, underscoring the effectiveness of iterative questioning in gathering comprehensive

Table 1. Overall performance in all GI cancers

Metrics	Simple	Templated	ICL	Multi-round
Understanding Medical History (0-20)	19.1±2.5	19.9±0.8	20.0±0.0	20.0±0.0
Diagnosis and Differential Diagnosis (0-20)	18.3±3.4	19.4±2.2	19.4±1.9	19.9±0.8
Further Examination and Reason (0-10)	6.6±3.8	8.8±2.6	9.0±2.5	9.8±1.1
Principles and Plans of Treatment (0-20)	10.8±4.0	14.0±3.3	14.2±4.7	16.5±3.2
Breadth and Depth of Results (0-20)	11.4±2.2	14.2±2.1	14.3±1.7	15.2±1.5
Thinking and Expressing Ability (0-10)	5.3±2.0	8.1±2.4	8.0±2.7	9.8±1.1
Total Score (0-100)	71.5±11.2	84.4±7.5	84.9±10.0	91.2±4.0

Table 2. Overall performance in gastric cancer

Metrics	Simple	Templated	ICL	Multi-round
Understanding Medical History (0-20)	18.1±3.3	19.8±1.1	20.0±0.0	20.0±0.0
Diagnosis and Differential Diagnosis (0-20)	16.4±4.1	18.8±3.0	18.8±2.6	19.8±1.1
Further Examination and Reason (0-10)	5.5±3.7	8.8±2.6	8.8±2.6	9.5±1.5
Principles and Plans of Treatment (0-20)	9.8±4.2	13.6±3.8	14.3±4.7	16.4±3.8
Breadth and Depth of Results (0-20)	11.0±2.0	15.0±1.5	14.5±1.5	15.7±1.7
Thinking and Expressing Ability (0-10)	5.0±1.5	9.0±2.0	8.6±2.3	9.5±1.5
Total Score (0-100)	65.7±10.6	85.0±7.7	85.0±10.9	91.0±5.0

Table 3. Overall performance in colorectal cancer

Metrics	Simple	Templated	ICL	Multi-round
Understanding Medical History (0-20)	20.0±0.0	20.0±0.0	20.0±0.0	20.0±0.0
Diagnosis and Differential Diagnosis (0-20)	20.0±0.0	20.0±0.0	20.0±0.0	20.0±0.0
Further Examination and Reason (0-10)	7.3±3.3	9.1±1.9	9.1±1.9	10.0±0.0
Principles and Plans of Treatment (0-20)	10.5±4.0	13.6±3.1	12.3±5.8	17.3±2.5
Breadth and Depth of Results (0-20)	11.4±2.2	14.1±1.9	14.1±1.9	15.0±0.0
Thinking and Expressing Ability (0-10)	4.5±1.4	7.3±2.5	6.8±3.2	10.0±0.0
Total Score (0-100)	73.6±6.4	84.1±6.3	82.3±10.7	92.3±2.5

Table 4. Overall performance in other GI cancer

Metrics	Simple	Templated	ICL	Multi-round
Understanding Medical History (0-20)	19.4±1.6	20.0±0.0	20.0±0.0	20.0±0.0
Diagnosis and Differential Diagnosis (0-20)	19.4±1.6	20.0±0.0	20.0±0.0	20.0±0.0
Further Examination and Reason (0-10)	7.6±3.9	9.1±2.6	9.4±2.4	10.0±0.0
Principles and Plans of Treatment (0-20)	11.2±4.0	14.7±3.2	15.6±3.4	15.6±3.4
Breadth and Depth of Results (0-20)	11.5±2.3	13.5±2.3	14.1±1.9	14.7±1.2
Thinking and Expressing Ability (0-10)	5.9±2.6	7.9±2.5	8.5±2.3	10.0±0.0
Total Score (0-100)	75.0±11.6	85.3±7.8	87.6±7.1	90.3±3.6

information. Second, in the context of 'Principles and Plans of Treatment', the marked improvement along the four types of prompts indicates the importance of diverse and complex prompts in formulating an effective treatment plan. The total scores also follow the same trend, with the Multi-strategy achieving the highest scores across all disease conditions consistently. These results provide strong evidence supporting the effectiveness of employing various prompt engineering strategies, as well as iterative questioning, in enhancing the performance of GPT-4 in medical contexts. However, the exact impact and effectiveness may vary depending on the specific disease condition, necessitating further nuanced analysis.

## DISCUSSION

To the best of our knowledge, it is evident that our research is pioneering in exploring the techniques to optimize LLMs specifically for the recommendation of treatments for gastrointestinal cancers. In contrast to studies that solely use simple prompts,<sup>2,5,6</sup> our research assessed a series of prompt engineering strategies including simple prompts, templated prompts, ICL, and multi-round interaction. Our results demonstrate that complex prompting approaches, especially multi-round interaction, are capable of accruing sufficient diagnostic and therapeutic information pertinent to a specific case. This approach facilitates the rational and efficient expansion of the storage of knowledge, thereby substantially enhancing model performance in collecting medical histories, forming accurate diagnoses, and recommending effective treatments for digestive cancers. The iterative nature of multi-round interaction consistently yielded the highest scores across evaluation metrics, highlighting its reliability and broad applicability. Our study also necessitates further exploration. Firstly, our adopted metric, based on clinicians' examinations, retains some degree of subjectivity. This accentuates the necessity for a more objective clinical evaluation method. We are currently collaborating with statisticians to devise novel evaluation tools to measure model performance more accurately and objectively. Moreover, our investigation was solely focused on tumors in the digestive system, indicating that future research could extend the application of LLMs to other types of cancers. Additionally, our primary clinical scenario was set in China, with the study conducted in Chinese before being translated into English using GPT-4. Although GPT-4 exhibits robust cross-language performance<sup>1</sup>, the influence of the selected language on performance deserves further study. Moreover, we have conducted preliminary assessments of various LLMs including GPT-4, Claude, ChatGLM, Wenxin Yiyao, and PaLM. Our findings indicate that GPT-4 exhibits superior underlying capabilities compared to its counterparts, leading us to select it as a representative of LLMs. Nevertheless, a comprehensive examination of the diverse and continually evolving LLMs is still an imperative area of future research. Last but not least, it is crucial to note that the data used to train GPT-4 predominantly originates from sources outside of China. However, due to variations in clinical guidelines, available medical technologies, perceptions of risk and benefit by patients and physicians, as well as disease prevalence trends in different regions, treatment approaches for gastrointestinal cancers can significantly differ across various regions. Subsequently, the outputs generated by GPT-4 may not entirely apply to Chinese patients. This particular aspect could potentially impact the evaluation scores during our comparative experiments of different prompt engineering strategies. To mitigate this potential issue, fine-tuning the LLMs with data specifically sourced from China could provide a more appropriate approach.

Recognizing the constraints and potential biases of LLMs is essential for their responsible and ethical application. One major concern is that LLMs gather knowledge from vast amounts of internet data that may contain inherent biases or inaccuracies. To mitigate potential bias and increase the reliability of our results, we employed a method of inter-rater reliability where each output from the model was independently evaluated by two separate individuals. Their evaluations were then compared and reconciled, ensuring a more objective and balanced assessment of the model's performance. Data privacy and security must be underscored when providing medical records to online LLMs. Thus, we have implemented stringent data protection measures, ensuring all patient data is anonymized and encrypted to protect privacy. Furthermore, the inadequate judgment and critical thinking skills of LLMs when interpreting medical records limit their performance in highly

specialized tasks. To address this, we've fostered close collaboration with expert clinicians and used prompt engineering to assist the model in understanding and handling complex medical information. We envision LLMs not as replacements for healthcare professionals, but rather as effective aid for clinical decision-making when properly guided. Future technological advancements, such as parameter-efficient fine-tuning for specialized tasks and the use of vectorized databases, may further contribute to solving these issues, offering better solutions for data security and private model deployment.

As we move forward, our findings open up avenues to further refine prompt engineering techniques to optimize LLMs for analyzing patient data and medical literature to recommend evidence-based treatments for digestive system cancers. We aim to explore how different prompts impact the model's ability to accurately recommend optimal interventions based on tumor characteristics and patient factors. For instance, certain prompts may enhance the model's capacity to suggest appropriate surgical procedures depending on tumor size, location, and staging. Other prompts could optimize the recommendation of systemic therapies like chemotherapy regimens and radiation therapy protocols tailored to the individual's medical history and cancer biomarkers. Advances in prompt engineering to account for all relevant clinical variables could enable the generation of more personalized and effective treatment plans for each unique patient case. However, more research is still urgently needed to ensure patient safety, avoid biases, and enable reliable interpretation of model outputs before these systems are ready for real-world clinical implementation. We must rigorously test prompts to identify any that skew recommendations in inappropriate or unsafe ways. Transparent reporting of model limitations and close collaboration with medical experts will be critical to responsible prompt engineering. While our results demonstrate immense promise for LLMs to enhance evidence-based decision support, translating these tools into practice will require thoughtful and ethical design paired with extensive validation to evolve prompt engineering strategies that provide trustworthy guidance without ever replacing human clinical judgment. Overall, steering LLMs through carefully crafted prompts shows great potential to augment clinicians' abilities to optimize and personalize treatment plans, propelling more effective cancer care. But as this technology continues maturing, maintaining patient well-being through rigorous prompt optimization and evaluation remains imperative.

## CONCLUSION

This study has underscored the potential and challenges associated with the application of prompt engineering techniques to large language models (LLMs) in the field of clinical oncology. Through careful crafting of simple, templated prompts and more complex strategies, like in-context learning (ICL) and multi-round interaction, we have seen promising capabilities of these models in processing and interpreting intricate medical data related to gastrointestinal cancers. This can substantially support healthcare professionals in making decisions about recommended treatments. However, it is crucial to continuously address the inherent limitations of these models, including potential biases, data privacy concerns, and their specific interpretative limitations in this clinical context. Although complex prompts, especially those allowing for iterative questioning, have shown great promise in optimizing the performance of LLMs, it's evident that further investigations are needed to refine these strategies and explore their potential integrations. As our study was conducted in a clearly defined and constrained environment to ensure consistency, further exploration in diverse settings is warranted to fully exploit the potential of LLMs in healthcare scenarios.

## REFERENCES

1. OpenAI, R. (2023). GPT-4 technical report. arXiv 2303.08774. DOI: 10.48550/arXiv.2303.08774.
2. Lee, P., Bubeck, S., and Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* **388**: 1233-1239. DOI: 10.1056/NEJMs2214184.
3. Lee, P., Goldberg, C., and Kohane, I. (2023). *The AI revolution in medicine: GPT-4 and beyond* (Pearson Education, Limited).
4. Xu, Y., Liu, X., Cao, X., et al. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* **2**: 100179. DOI: 10.1016/j.xinn.2021.100179.
5. Nori, H., King, N., McKinney, S.M., et al. (2023). Capabilities of GPT-4 on medical

- challenge problems. arXiv preprint arXiv:2303.13375. DOI: 10.48550/arXiv.2303.13375.
6. Ayers, J.W., Poliak, A., Dredze, M., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**: 589-596. DOI: 10.1001/jamainternmed.2023.1838.
  7. Haver, H.L., Ambinder, E.B., Bahl, M., et al. (2023). Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* **307**: e230424. DOI: 10.1148/radiol.230424.
  8. Zhu, L., Mou, W., and Chen, R. (2023). Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J. Transl. Med.* **21**: 1-4. DOI: 10.1186/s12967-022-03835-4.
  9. Uprety, D., Zhu, D., and West, H.J. (2023). ChatGPT—a promising generative AI tool and its implications for cancer care. *Cancer* **129**: 2284-2289. DOI: 10.1002/cncr.34827.
  10. Zhong, Y., Chen, Y.J., Zhou, Y., et al. (2023). The artificial intelligence large language models and neuropsychiatry practice and research ethic. *Asian J. Psychiatr.* **84**: 103577. DOI: 10.1016/j.ajp.2023.103577.
  11. Young, J.N., Ross, O.H., Poplousky, D., et al. (2023). The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. *J. Am. Acad. Dermatol.* **89**: 602-604. DOI: 10.1016/j.jaad.2023.05.024.
  12. Xie, Y., Seth, I., Hunter-Smith, D.J., et al. (2023). Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesth. Plast. Surg.* **47**: 1985-1993. DOI: 10.1007/s00266-023-03338-7.
  13. Buzzaccarini, G., Degliuomini, R.S., and Borin, M. (2023). The artificial intelligence application in aesthetic medicine: How ChatGPT can revolutionize the aesthetic world. *Aesth. Plast. Surg.* **47**: 2211-2212. DOI: 10.1007/s00266-023-03416-w.
  14. Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog* **1**: 9. [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf).
  15. Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *NeurIPS* **33**: 1877-1901. DOI: 10.48550/arXiv.2005.14165.
  16. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS* **33**: 9459-9474. DOI: 10.48550/arXiv.2005.11401.
  17. Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS* **35**: 24824-24837. DOI: 10.48550/arXiv.2201.11903.
  18. Zhou, D., Schärli, N., Hou, L., et al. (2022). Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625. DOI: 10.48550/arXiv.2205.10625.
  19. Yao, S., Yu, D., Zhao, J., et al. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601. DOI: 10.48550/arXiv.2305.10601.
  20. Fu, Y., Peng, H., Sabharwal, A., et al. (2022). Complexity-based prompting for multi-step reasoning. arXiv preprint arXiv:2210.00720. DOI: 10.48550/arXiv.2210.00720.
  21. Khot, T., Trivedi, H., Finlayson, M., et al. (2022). Decomposed prompting: A modular approach for solving complex tasks. arXiv preprint arXiv:2210.02406. DOI: 10.48550/arXiv.2210.02406.
  22. White, J., Fu, Q., Hays, S., et al. (2023). A prompt pattern catalog to enhance prompt

engineering with ChatGPT. arXiv preprint arXiv:2302.11382. DOI: 10.48550/arXiv.2302.11382.

23. Suzgun, M., Scales, N., Schärli, N., et al. (2022). Challenging big-bench tasks and whether chain-of-thought can solve them. arXiv preprint arXiv:2210.09261. DOI: 10.48550/arXiv.2210.09261.

## FUNDING AND ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (91959205 to L.S., U22A20327 to L.S., 82203881 to Y.C., 82272627 to X.T.Z., 7232018 to Y.S., 12090022 to B.D., 11831002 to B.D., 81801778 to L.Z.), Beijing Natural Science Foundation (7222021 to Y.C., Z200015 to X.T.Z.), Beijing Hospitals Authority Youth Programme (QML20231115 to Y.C.), Clinical Medicine Plus X-Young Scholars Project of Peking University (PKU2023LCXQ041 to Y.C. and L.Z.). Guangdong Provincial Key Laboratory of Precision Medicine for Gastrointestinal Cancer (2020B121201004). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## AUTHOR CONTRIBUTIONS

J. Yuan, P. Bao, Z. Chen, M. Yuan, and J. Pan contributed to data analysis and interpretation, and drafted the manuscript. J. Zhao, Y. Xie, Y. Cao, Y. Wang, Z. Wang, Z. Lu, X. Zhang, J. Li and L. Ma performed the sample preparation. Y. Chen, L. Zhang, L. Shen, and B. Dong planned the study and participated in manuscript revision. All authors have given final approval for the manuscript to be published and have agreed to be responsible for all aspects of the manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## ETHICAL STATEMENT AND PATIENT CONSENT

Not applicable.

## DATA AND CODE AVAILABILITY

A pre-trained Chinese BERT model in Hugging Face (<https://huggingface.co/hfl/chinese-bert-wwm-ext>), specifically the "hfl/chinese-bert-wwm-ext", is utilized to translate patient conditions into a high-dimensional embedding space (768 dimensions in this study), capturing the context of the condition effectively.

## SUPPLEMENTAL INFORMATION

It can be found online at <https://doi.org/10.59717/j.xinn-med.2023.100019>

## LEAD CONTACT WEBSITE

<http://faculty.bicmr.pku.edu.cn/~dongbin/>