scientific reports

Check for updates

OPEN

LLM-based medical dialogue dataset generation with automated instructions

Hao Zhou[1,2,4]✉, Xuanlang Hu[1,4], Ting He[1]✉, Haibin You[1], Jiyu Yin[1], Jingjing Xu[1], Zhizhong Lin[1] & Huazhen Wang[1,3]

Constructing medical dialogue datasets poses significant challenges owing to legal and privacy concerns. In the wake of the advancement of large language models (LLMs), automated instruction generation grounded in LLMs has emerged as a promising approach for dataset construction. However, the existing methods often overlook the integration of domain knowledge, such as the standards and regulations stipulated in official documents. This renders the generated instructions and corpus of reduced value. In this work, we propose a new LLM-based automated instruction generation framework to build a medical dialogue dataset compliant with the guidelines of Medical Chinese Test (MCT). The framework involves the construction of a hand-crafted instruction set, corpus refinement, instruction sampling using maximum marginal relevance, and the K-means algorithm. By incorporating domain-specific knowledge and adopting instruction sampling strategy, the generated instructions and corpus basically meet the MCT standards. We tested this generation framework in the experiment with ChatGPT (gpt-3.5-turbo) and the medical LLM model Zuoyi, finding that compared to real-world medical dialogue datasets, the generated dataset MCT-Chat consisting of 20k examples demonstrates excellent performance in terms of both objective and subjective indicators.

Medical Chinese Test (MCT) (https://www.mandarin.ac.cn/tests/mct-medical-chinese-test.html) is a standardized assessment designed to evaluate the Chinese language proficiency of non-native speakers in the medical domain. It places significant emphasis on practical tasks that require students to communicate with patients, medical staff, and other relevant individuals in daily hospital settings using the Chinese language. Consequently, it is highly reliant on extensive multi-turn dialogue corpora that are abundant in medical scenarios and adhere to medical principles. However, due to patient privacy concerns, the collection and annotation of medical dialogue data is an extremely challenging task. Additionally, compared to open-domain dialogue datasets, medical dialogue data demand a much higher level of professionalism[1]. First, medical dialogues occur in specific medical contexts, characterized by a large number of dialogue turns and strong contextual coherence. For example, a doctor's diagnosis and treatment process may involve multiple rounds of questions and answers with a patient to fully understand the symptoms and medical history. Second, medical dialogues have strict role limitations, requiring adherence to the language habits and emotional expressions of doctors and patients. For instance, doctors usually use professional medical terms, while patients tend to use more common and descriptive language. Moreover, medical dialogues should incorporate vertical domain knowledge and must meet the requirements of guiding documents[2]. As a result, constructing an appropriate medical dialogue dataset for MCT presents significant difficulties.

With the rapid development of large language models (LLMs), it has been successfully applied to a variety of areas. In medicine-related areas, LLMs has shown great potential in clinical decision systems[3], medical image analysis[4], electronic health record processing[5], medical education and training[6]. LLMs-based generation of high-quality medical datasets represents one of the most promising applications of large language models in healthcare, addressing critical challenges related to data scarcity, privacy constraints, and the need for topic diversity[7,8]. A typical construction method based on LLMs depends on high-quality fine-tuning instructions, which are clear and standardized guidelines to fulfill specific tasks or goals[9]. Appropriately designed fine-tuning instructions can remarkably improve the generation quality in low-resource domains.

Currently, LLM-based data generation heavily depends on manual instructions, requiring extensive human effort for fine-tuning to create a good instruction dataset, as exemplified by Prompsource[10] and

[1]School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China. [2]Xiamen Key Laboratory of Data Security, Huaqiao University, Xiamen 361021, China. [3]Provincial Key Laboratory of Computer Vision and Machine Learning, Huaqiao University, Xiamen 361021, China. [4]Hao Zhou and Xuanlang Hu contributed equally to this work. ✉email: haozhou@hqu.edu.cn; xuantinghe@hit.edu.cn

SuperNaturalInstructions[11]. Manually constructing instructions not only takes time and effort but can also be influenced by human subjectivity and prone to errors. To address these limitations, the automated instruction generation framework has attracted growing interest[12–14]. However, these studies exhibit certain limitations. First, the automated instruction generation is typically updated iteratively based on model feedback, lacking the integration of domain knowledge and making generated instructions less interpretable. Second, using model-generated results as an evaluation criterion overlooks the direct assessment of the generated instructions. Thus, the instructions and corpus generated for MCT through these methods cannot meet the desired standards.

In this work, we propose a new LLM-based Automated Instruction Generation framework for MCT multi-turn dialogue (AIG-MCT). AIG-MCT starts by constructing a manual instruction set with a subject description set and sample list as the instruction pool's initialization. Instructions from the pool are fed into an LLM, generating machine instructions and multi-turn dialogues. The maximum margin relevance (MMR) algorithm and K-means are then used for machine instruction updating in the instruction pool, and through this iterative process, a multi-turn MCT dialogue corpus is created based on automated instructions. By incorporating domain-knowledge and instruction sampling strategy, the MCT corpus generated through this framework demonstrates outstanding performance, as tested in the experiment with ChatGPT (gpt-3.5-turbo) and the medical large language model ZuoYi. The major contributions of this work are summarized as follows:

- A LLM-based automated instruction generation framework is proposed to construct medical dialogue datasets that meet the specified standards and guidelines in MCT.
- The framework features the adoption of domain-specific knowledge and instruction sampling strategy, and is used to generate a vertical-domain dataset comparable to real-world datasets.
- The generated dataset MCT-Chat consisting of 20K medical dialogues and the corresponding instructions are released and made accessible for academic research.

Related work

In the medical field, large language models as a transformative approach has been widely used to facilitate many applications[12]. For example, the integration of LLMs with biomedical knowledge graphs can combine the reasoning capabilities of LLMs with the structured knowledge representation of graphs, achieving promising results in drug-drug interactions and drug discovery[15]. LLMs-based data generation has attracted great attention thanks to the powerful generative ability of LLMs. Typical works based on GPT include[16], which developed a dataset to train a consultation model;[17], which used GPT to generate improved suggestions for alerts in clinical decision support systems;[18], which focused on the text generation in healthcare domain. In particular, Li et al.[19] used ChatGPT to generate a doctor-patient dialogue corpus based on a database of approximately 700 diseases and their related symptoms, medical tests, and recommended medications. Das et al.[20] generated patient-physician dialogue from clinical notes using a single LLM. Their works demonstrate the feasibility and efficiency of LLM-based medical dialogue corpus generation.

The key to generating data using LLMs is automated generation of fine-tuning instructions. Generally, there are three categories of instruction generation methods, based on templates, chain-of-thought, or iterative learning. Instruction generation based on templates or rules is straightforward, as shown in[11], which proposed Super-NaturalInstructions to generate simple instruction templates. It cannot deal with complex medical scenarios and is more suitable to simple tasks. Chain-of-thought generation methods exhibit strong creativity. For example, Liu et al.[21] proposed Think-Then-Write, which reassembled the generated questions and answers into a chain of thought as instructions. However, this approach has difficulty controlling the length and complexity, and it requires excessive computational resources and time. Instruction generation based on iterative learning leverages the feedback of the LLM itself to continuously optimize instructions. By using historical input-output data and error information, it refines and optimizes control instructions, making it a relatively superior and efficient method. For instance, Wang et al.[13] proposed Self-Instruct, which generates a large number of instructions, inputs, and output samples from the language model itself, screens and revises them, and then uses them to fine-tune the original language model. However, iterative learning-based methods rely on the model's self-generation capability and feedback, which may result in low instruction diversity and accuracy during the generation process. Moreover, iteratively fine-tuning the language model can be cumbersome and time-consuming.

Despite these achievements, few works involve the generation of domain-specific medical dialogue corpus that require compliance with explicit standards and regulations. Compared to universal medical dialogue corpora, the MCT corpus must meet vertical knowledge constraints, including the task outline and topic outline, to ensure the interpretability of the generated instructions and the standardization of the generated corpus. Following the method of iterative learning-based instruction generation, we incorporate domain-knowledge to ensure the feasibility of the resulted corpora, and moreover, adopt instruction sampling during the generation process to address the issue of low instruction diversity.

Method

In order to construct MCT multi-turn dialogue corpus through automated instruction generation, we present the framework AIG-MCT (short for Automated Instruction Generation for Medical Chinese Test). As illustrated in Fig. 1, AIG-MCT encompasses four key stages:

(1) Construction of the manual instruction set;
(2) Instruction generation and corpus correction;
(3) Instruction sampling based on the MMR algorithm;
(4) Instruction pool update based on the K-Means algorithm.

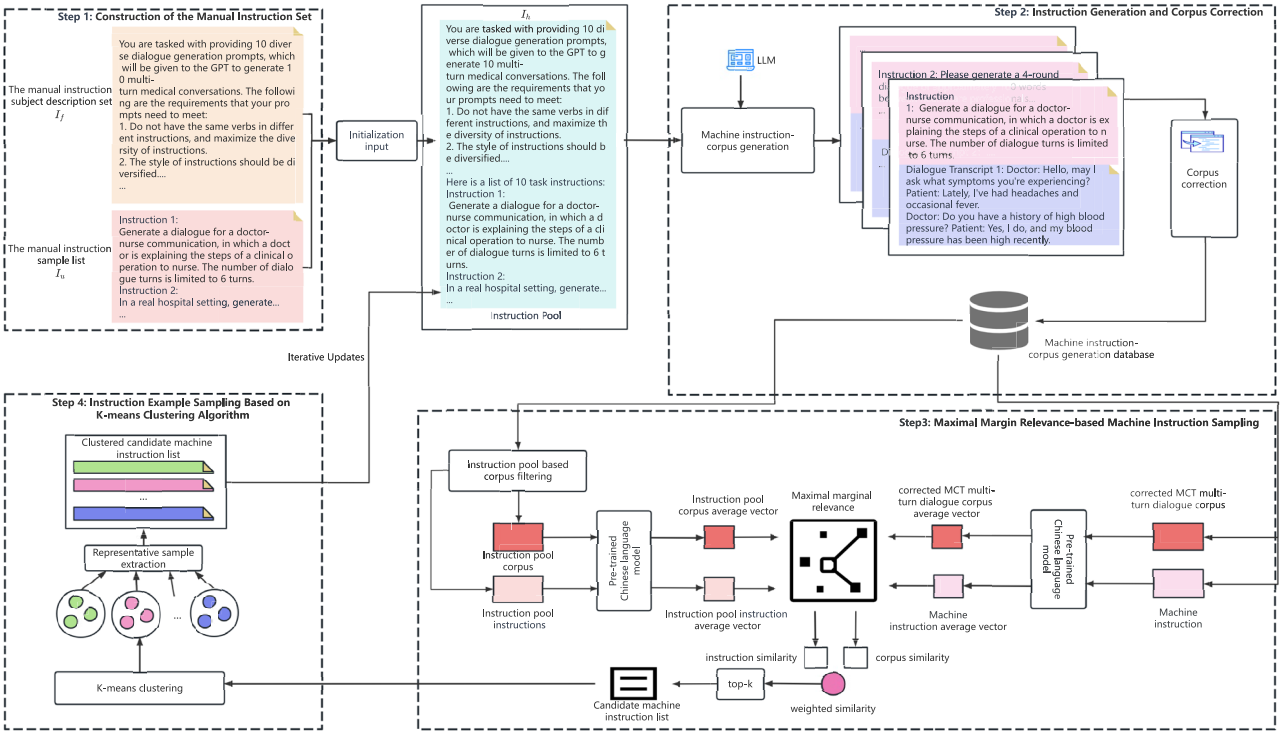

Fig. 1. The entire framework consists of four stages. First, the hand-crafted instruction subject description set I_f and example list I_u are provided for initialization of the instruction pool. When they are fed into LLM, multiple copies of instructions and corresponding corpora are obtained (the generated corpora will be corrected for enhancement), forming the database which comprises both machine instructions and generated corpora. To effectively update the instruction pool, in Step 3, MMR is used to select those most distinct instruction candidates based on the cosine similarity between the instructions as well as between the corpora in the database, and finally, K-means is employed to identify the most representative instructions for updating.

Construction of the manual instruction set

The instruction set functions as a set of exemplars to guide the large language model (LLM) in generating innovative and diverse instructions. We first construct a manual instruction set $I_h = \{I_f, I_u\}$, where I_f is the instruction subject description set, and I_u is the instruction sample list. $I_f = \{x_1, ..., x_n\}$, where x_i corresponds to a distinct description of restrictions. These descriptions are derived from domain-specific knowledge or constraints, such as the content, style, type, length, and turns of dialogue. The combined constraint description constructs an instruction sample list I_u. Here is an example of I_u:

> Please provide a dialogue between a pediatrician and parents, in which the pediatrician queries the feeding habits of the infant. The dialogue should consist of 3 turns, each turn comprising approximately 40 words. Ensure that the multi-turn dialogue aligns with the context of a pediatric outpatient scenario.

The manual instruction set I_h is used for initialization of the instruction pool I_{pool}, which serves as the input to the LLM. Note that I_u will be iteratively updated with machine instructions in the subsequent steps, and we use I_s to replace I_u in the instruction pool, i.e., $I_{pool} = \{I_f, I_s\}$.

Instruction generation and corpus correction

The LLM utilizes the instructions in the pool to generate both the dialogue corpus and machine instructions. The output can be denoted as (i_m, d), where i_m represents the machine instruction and d represents the corresponding corpus. Due to the uncontrollability of LLMs, the generated corpus may include not only dialogue text but also other undesired formatted data. In addition, there may be too many roles in the dialogue beyond the scope defined in MCT. To address these issues, post-processing on the generated examples is necessary. As illustrated in Fig. 2, format filtering is first conducted to remove non-dialogue data, including JSON-formatted data, patient questions, and electronic medical records. Then, role and relationship standardization is executed, consisting of three steps: identify roles, extract unique roles, and map the roles to the specified relationship types in MCT (doctor-patient, doctor-medical staff, doctor-doctor, and patient-medical staff). Finally, to further enhance the corpus, the following indicators are used as thresholds to filter the examples:

- Medical vocabulary coverage: A minimum coverage threshold g_c is set to evaluate if the generated corpus cover an adequate range of medical vocabulary in the MCT outline. Specifically, we match the result of word

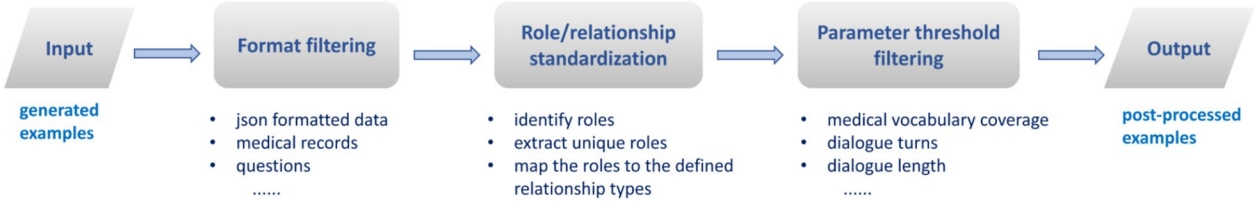

Fig. 2. Post-processing on the generated examples.

segmentation with the official medical vocabulary to compute the coverage and if it does not reach g_c, we will drop the corpus along with its corresponding instruction.
- Number of dialogue turns: For medical dialogue generation tasks, it is usually necessary to simulate multi-turn dialogue interactions that occur in real healthcare scenarios. Thus, a minimum turn threshold g_t is established. The number of dialogue turns (by counting punctuation marks) in the corpus is compared with g_t. If it does not reach g_t, the corpus and its corresponding instruction are dropped.
- Dialogue length: Minimum length threshold g_{min} and maximum length threshold g_{max} are defined to ensure that the dialogue length in the corpus falls within the desired range. Drop the corpus and instruction if the length of the dialogue is not in the range $[g_{min}, g_{max}]$.
- Grammar correction (optional): We use HanLP's grammar analysis tool to correct grammar errors in the dialogue corpus, such as incomplete sentences and misspelled words, to ensure the grammatical accuracy of the generated corpus.

The enhanced MCT multi-turn dialogue corpus is denoted as d_e. After multiple iterations, we have the database $D_e = \{(i_m, d_e)\}$, and in particular, the corpus generated using the instruction pool (representing the exemplar instructions) is denoted as D_s, i.e., $D_s = \{(i_s, d_s)\}$, where i_s represents one instruction from I_s and d_s represents its generated corpus.

MMR-based instruction sampling

In order to increase the difference between the exemplar instructions in I_s and the instructions in D_e so that the generated examples demonstrate higher diversity, we utilize the MMR algorithm to sample the machine instructions in D_e. This algorithm first uses cosine similarity to measure the similarity between i_m and i_s ($m \neq s$). It also measures the similarity between the corpus d_e and d_s. The marginal distance is then calculated through a weighted summation, and suitable machine instructions are selected based on this criterion.

Initially, the pre-trained large language model (LM) is used to represent the data as vectors. This results in the vector representation of the database as $V_{D_e} = (V_{i_m}, V_{d_e})$. The machine instruction vector V_{i_m} is obtained by performing average pooling on the word vectors in i_m, i.e.,

$$V_{i_m} = \frac{1}{N} \sum_{i=1}^{N} LLM(w_i) \tag{1}$$

where w_i represents the *i*-th word in the machine instruction i_m, and *N* is the number of words. The vector V_{d_e} is obtained by performing average pooling on the word vectors in the corpus d_e, calculated as

$$V_{d_e} = \frac{1}{M} \sum_{i=1}^{M} LLM(z_i) \tag{2}$$

where z_i represents the *i*-th word in d_e, and *M* denotes the number of words. We obtain V_{I_s} and V_{d_s} in a similar way.

Next, the cosine similarity is calculated between V_{i_m} and V_{I_s}, as well as between V_{d_s} and V_{d_e}, producing S_{ins} and S_d, respectively. S_{ins} and S_d are combined through a weighted summation to obtain the weighted similarity S_c, as

$$S_c = \delta \cdot S_{ins} + (1 - \delta) \cdot S_d \tag{3}$$

where δ represents the factor used to control the contribution ratio between instruction similarity and corpus similarity.

For instruction i_s, the top-*k* machine instructions from D_e that produce the smallest S_c are selected as candidate instructions for the update of the instruction pool. All the candidate instructions form a list, denoted as T_m.

Update instruction pool with K-means algorithm

We apply the *K*-means algorithm to cluster the machine instructions in the list T_m. The objective of the *K*-means algorithm is to divide the candidate machine instructions into *K* clusters, where *K* is equal to the size of the

instruction sample list in the instruction pool. In order to estimate the distance between each instruction and its corresponding cluster centroid, we define the objective function *J* as

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \tag{4}$$

where C_i represents the *i*-th cluster, and μ_i represents the centroid of C_i. By minimizing *J*, the centroid of each cluster is obtained. The candidate instruction closest to μ_i is identified as x_{μ_i}, which is the representative instruction in C_i, and is added to the updated machine instruction list T_K.

To gradually reduce the influence of human instruction during the automated instruction generation process, we progressively reduce the use of human instruction examples while simultaneously increasing the utilization of machine instructions. Specifically, we introduce a decay rate parameter α that varies between 0 and 1 to control the number of instructions that need to be updated from I_s. A certain number of instructions are randomly removed from I_s, and an equal number of instructions are randomly selected from T_K for complement. The number for the *n*-th round updating is given as

$$N_{u_n} = (1 - \alpha^n) * N_{I_{s_{n-1}}} \tag{5}$$

where $N_{I_{s_{n-1}}}$ represents the size of the instruction list I_s in the $(n-1)$-th turn, and N_{u_n} denotes the number of instructions to be updated in the *n*-th turn.

Experiment

In order to validate the effectiveness of the proposed framework, we employed ChatGPT (gpt-3.5-turbo) as the basic generative model, and the medical large language model ZuoYi to check and refine the generated corpus. Human expert validation was conducted at the last step to ensure the medical accuracy and the consistency to the requirements of Medical Chinese Test (MCT).

Setting domain-knowledge constrains

The domain-specific knowledge constraints are sourced from the MCT outline. These constraints encompass medical vocabulary guidelines, task guidelines, topic guidelines, dialogue length specifications, the required number of dialogue turns, and dialogue scenarios. As presented in Table 1, the medical vocabulary guidelines pertain to the vocabulary dictionary defined by MCT. The task guidelines include the list of clinical tasks, and the topic guidelines comprise the list of clinical topics specified in MCT. The dialogue length related to the "Chinese Proficiency Grading Standards for International Chinese Language Education" at the Level 4 is empirically set less than 500. The number of dialogue turns specifies the minimum number required for a complete medical dialogue and is set to be greater than or equal to 2 (the choice of this value are verified in the ablation study). The dialogue scenarios refer to the department of treatment where the medical dialogue takes place.

Manual instruction set construction

The manual instruction set $I_h = \{I_f, I_u\}$ is established. The instruction subject description set I_f includes task goal definitions, specifications for instruction targeting and dialogue generation, such as dialogue length, number of dialogue turns, and diversification of instruction. The task goal definition enables the LLM to generate medical dialogues through instructions. The specifications ensure that LLM generates diverse and adaptable instructions that align with MCT outline. The instruction sample list I_u is created by combining the factors defined in I_f to form 10 instruction samples. The instruction subject description set I_f in the experiment is designed as follows:

(1) You are tasked with providing 10 diverse dialogue generation prompts, which will be given to the GPT to generate 10 multi-turn medical conversations. The following are the requirements that your prompts need to meet:
(2) Do not have the same verbs in different instructions, and maximize the diversity of instructions.

Item	Description
Task guidelines	Medical history collection, physical examination, disease diagnosis, clinical operation, discussion of treatment plans, diagnosis & treatment plans formulation, disease prevention, interpersonal
Topic guidelines	D–D inquiry, D–D communication, D–D instruction, D–P consultation,D-P diagnosis, D–P treatment, D–P complaint, D–P communication, D–p instruction, D–N communication, D–N instruction, P–N communication, P–N instruction, P–N consultation(D for doctor, P for patient, and N for nurse)
Medical vocabulary guidelines	Amoebiasis, aspirin, canceration, cancer, AIDS, belching, lips, follow, routine, sitting height, ... (A total of 1500 medical words)
Dialogue length	< 500
Dialogue turns	≥ 2
Dialogue scenarios	Emergency room, operation room, etc.

Table 1. Domain-knowledge constrains.

(3) The style of instructions should be diversified. For example, the doctor's conversation style includes extreme expression, soothing expression, authoritative expression, question options, additional question expression, etc.
(4) The emotions of the instructions are diverse. For example, the patient's conversational emotion should include positive, negative, and neutral, while the doctor's conversational emotion is neutral.
(5) Instructions are written in Chinese.
(6) You should randomly select words from the MCT task guidelines and MCT topic guidelines to generate instructions. MCT task guidelines: medical history collection, physical examination, disease diagnosis, clinical operation, treatment plan discussion, diagnosis and treatment plan formulation, disease prevention, and interpersonal communication. MCT topic guidelines: doctor–doctor inquiry, doctor–doctor communication, doctor–doctor instruction, doctor–patient consultation, doctor–patient diagnosis, doctor–patient treatment, doctor–patient complaint, doctor–patient communication, doctor–patient instruction, doctor–nurse communication, doctor–patient instruction, patient–nurse communication , patient-nurse instruction, patient–nurse consultation.
(7) Instructions should contain a variety of dialogue scenarios, which are required to be close to real medical scenarios.
(8) Instructions should include the number of dialogue turns, and the minimum turn should not be smaller than 2.
(9) Instructions should contains 1 or 2 sentences.
(10) Imperative or question sentences are allowed.

Two examples of I_u are given below.

- Create a dialogue for a doctor-patient diagnosis, where the doctor tells the patient their diagnosis and explains the possible causes of the disease. The number of dialogue turns is 4. The instruction contains 20 words. (Component 6, 8, and 9)
- You are a doctor, please inform the patient of the treatment plan for the disease in Chinese, the topic comes from the formulation of the diagnosis and treatment plan in the MCT task guidelines, and the dialogue is limited to 8 turns. (Component 5, 6, and 8)

More examples can be viewed in the Appendix.

Corpus correction and updating

All the generated examples from each iteration went through post-processing as illustrated in Fig. 2. Regarding parameter threshold filtering, the major parameters in the framework were set as in Table 2. Note that most of the parameter setting were referenced with the official document "Chinese Proficiency Grading Standards for International Chinese Language Education" at Level 4, which is highly popular among potential examinees in MCT. And all the used parameters were finally validated by Chinese education experts. Specifically, g_c was set 1 to ensure the generated data size, g_t was set 2 as the ablation studies in Section 4.7 verified. In choosing top-k machine instructions as candidates, k was empirically set 0.8, i.e., preserving 80% most similar examples. For K-means clustering, K was equal to the size of the instruction sample list in the instruction pool and that size was dynamically updated as shown in Eq. (5). By setting the initial instruction sample list I_u 10, about 500 dialogues can be produced after iterative-updating and post-processing. After 40 rounds of experiments (each with unique randomly created instruction sample list) and human screening, we ultimately built the dataset MCT-Chat consisting of 20k high-quality MCT multi-turn dialogues. In spite of human validation during the dataset construction, it should be emphasized that the generated corpus only serve as a potential resource for Medical Chinese Test, and any selected materials for real test must undergo rigorous manual estimation and validation to ensure the standardization of the test. A sample of MCT-Chat (translated in English) consisting of 1000 examples can be accessed via https://github.com/haozhou2018/dataset-mctchat. Please note that this dataset serves as potential materials for Medical Chinese Test, thus original dialogue examples and instructions are in Chinese.

Parameter	Description	Value
g_c	Minimal medical vocabulary coverage	1
g_t	Minimal turn threshold	2
g_{min}	Minimal length threshold	0
g_{max}	Maximum length threshold	500
δ	Weighting factor in Eq. (3)	0.8
k	Top-k similar examples	0.80
K	Clusters in K-means	adapting
α	Decay rate in Eq. (5)	0.9

Table 2. Major parameters in the experiment.

Assessment indicators

Various objective and subjective indicators are used to assess the generated dataset MCT-Chat. Regarding objective indicators, both linguistic diversity and content diversity are crucial. We employ two linguistic diversity indicators, Dt_1 and Dt_2, which involve calculating the proportions of different n-gram types in the data. In detail,

$$Dt_1 = \frac{N_{d\_gram1}}{N_{t\_gram1}} \tag{6}$$

where N_{d_gram1} represents the count of unique 1-grams, N_{t_gram1} refers to the total count of 1-grams. Dt_2 is calculated in a similar way. We evaluate content diversity through Disease Diversity (DD) and Symptom Diversity (SD). Formally,

$$DD = \frac{N_{dis\_types}}{N_{dis\_words}} \tag{7}$$

where N_{dis_types} indicates the number of disease types, N_{dis_words} refers to the total number of disease-related terms. Similarly,

$$SD = \frac{N_{sym\_types}}{N_{sym\_words}} \tag{8}$$

where the subscript *sym* stands for symptom. To assess the alignment between MCT-Chat and the MCT outline, we adopt the following objective indicators:

- Average coverage of medical vocabulary
- Average number of turns
- Average number of words
- Distribution of dialogue roles
- Topic types
- Task types

Regarding subjective indicators, the following aspects are considered for human evaluation:

- Fluency (*Flue*): Assessing whether the dialogue is smooth, natural, easily understandable, and whether the language usage is idiomatic.
- Tendency (*Tend*): Determining if the dialogue exhibits bias towards a specific answer or solution.
- Rationality (*Rati*): Evaluating whether the dialogue aligns with common medical knowledge and practical scenarios.
- Matching degree (*Matc*): Checking whether the dialogue meets the MCT outline.
- Discrimination in grading (*Disc*): Comparing the difficulty level with the designated MCT proficiency levels.

We randomly selected 100 dialogues from the dataset and presented them to five medical experts for evaluation. Each expert independently rated each example on a scale of 1 to 5, where 1 indicates poor quality and 5 indicates excellent quality. The expert ratings were averaged to obtain the final score. As part of the data estimation, we confirm that all the experiments in this study were carried out in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki. The experimental protocols were approved by the Ethics Review Committee of Huaqiao University Medical College. Informed consent was obtained from all individual participants included in the study.

Comparative datesets

We compared MCT-Chat and other typical medical dialogue datasets including MedDialog[1], MedDG[22], and DISC-Med-SFT[23]. The MedDialog dataset has both English version (MedDialog-EN) and Chinese version (MedDialog-CN). We selected the Chinese version, which contains approximately 1.1 million dialogues and 4 million utterances. We randomly sampled around 20K dialogues, covering 29 broad professional categories and 172 fine-grained professional categories, demonstrating a significant advantage in terms of dialogue volume and disease variety. The MedDG dataset comprises 17,864 Chinese dialogues, covering five major categories of medical entities: diseases, symptoms, medications, examinations, and attributes. The DISC-Med-SFT dataset, released by the Data Intelligence and Social Computing Laboratory at Fudan University, is a high-quality supervised fine-tuned dataset containing 470K dialogue examples. We randomly selected around 20K dialogues, covering various scenarios such as single-turn medical Q&A, multi-turn medical consultation, and medical multiple-choice questions. Regarding the human experts estimation, random 100 dialogues were selected from each dataset as the practice for Chat-MCT.

General results analysis

Table 3 reports the objective indicators on MCT-Chat and the comparative datasets. In terms of Dt_1 and Dt_2, it reveals that the MCT-Chat dataset outperforms others, demonstrating excellent lexical diversity and richness. With respect to *DD* and *SD*, we referenced the International Classification of Diseases (10th Revision, ICD-10) from the National Center for Health Statistics, and the Chinese Medical Knowledge Graph (CMeKG 2.0)

Dataset	Dt_1	Dt_2	*DD*	*SD*
MedDialog	0.396	0.752	0.075	0.040
MedDG	0.438	0.738	0.013	0.013
DISC-Med-SFT	0.364	0.705	0.073	0.041
MCT-Chat	0.443	0.760	0.041	0.031

Table 3. Objective assessment.

Dataset	Med. Vocab. coverage	Average turns	Average words	Samples	Role[1]	Topic[2]	Task
MCT-Chat	73.53%	9.6	382	20358	D–D; D–P; D–N; P-N	D–D(iq); D–D(cm); D–D (is); D–P(cs); D–P(d); D–P (t); D–P(cp); D–P(cm);D–P (is); D–N(cm);D–N(is); N–P(cm); P–N(is); P–N(iq)	medical history collection; physical examination; disease diagnosis; clinical operation; treatment plan discussion; diagnosis & treatment plan formulation; disease prevention; interpersonal communication
MedDialog	78.27%	8.6	449	20000	D–P	–	–
MEdDG	52.87%	11.4	254	17864	D–P	–	–
DISC-Med-SFT	77.13%	7.1	597	20000	D–P	–	–

Table 4. Consistency indicators with medical dialogue datasets. [1]Role: D for doctor, P for patient, and N for nurse. [2]Topic: iq for inquiry, cm for communication, is for instruction; cs for consultation; d for diagnosis, t for treatment, and cp for complaint

Dataset	*Flue*	*Tend*	*Rati*	*Matc*	*Disc*
MedDialog	4.8	1.9	4.7	4.6	4.5
MedDG	4.7	2.0	4.6	4.3	4.2
DISC-Med-SFT	4.8	1.8	4.8	4.7	4.6
MCT-Chat	4.8	1.8	4.9	4.7	4.8

Table 5. Human evaluation results.

for calculation. From Table 3, it is viewed that MCT-Chat exhibits 0.041 for *DD* and 0.031 for *SD*, which are comparatively lower than those of the MedDialog and DISC-Med-SFT datasets. It can be understandable that MCT-Chat was originally designed for examination and the purpose of MCT-Chat is to provide matching dialogues rather than cover all diseases and symptoms within the medical field. More objective indicators are presented in Table 4. As it shows, MCT-Chat exhibits satisfactory performance in medical vocabulary coverage, with its diverse dialogue roles, topic types and tasks, highlighting its advantage in simulating authentic medical scenarios. The average number of turns and word count suggest that the dialogues within this dataset are appropriate to assess the examinees' medical knowledge and communication skills, aligning with the MCT outline requirements.

With respect to human expert estimation, Table 5 presents human evaluation results across five metrics : *Flue* (Fluency), *Tend* (Tendency), *Rati* (Rationality), *Matc* (Matching degree), and *Disc* (Discrimination in grading). Generally, MCT-Chat achieves scores comparable to other real-world datasets, with near-identical performance in *Flue* (4.8) and *Matc* (4.7), and marginally higher scores in *Rati* (4.9) and *Disc* (4.8). The results confirm that MCT-Chat demonstrates satisfactory quality, making it potentially appropriate for official examination and other purposes of use.

It is noted that while our research demonstrates promising results in automated instruction generation for MCT multi-turn dialogues, limitations persist in semantic and contextual understanding. LLMs may misinterpret dialogue topics, leading to inaccuracies in complex medical instructions. Thus, manual review and refinement of generated content remain essential before real-world deployment. Furthermore, despite the comprehensive evaluation methods, potential biases can arise in subjective assessments.

Ablation study and discussion

For the parameters summarized in Table 2, besides those empirically set values, we particularly conducted ablation experiments with respect to the parameter K in K-means clustering and g_t (minimal dialogue turn threshold). In the original framework, K is set equal to the size of the instruction sample list in the instruction

Strategy	Dt_1	Dt_2	*DD*	*SD*	Final Dialogues
Self-adapting K	0.031	0.192	0.154	0.036	490
K = 5	0.268	0.673	0.539	0.322	10
K = 15	0.172	0.562	0.311	0.192	30

Table 6. Comparison of K setting strategies in K-means.

Value	Total dialogues	Dt_1	Dt_2	*DD*	*SD*	Average turns
$g_t = 1$	374	0.031	0.209	0.027	0.037	13.2
$g_t = 2$	490	0.031	0.192	0.154	0.036	10.4
$g_t = 4$	95	0.094	0.429	0.095	0.120	10.1

Table 7. Comparison of minimal dialogue turns threshold.

pool and thus dynamically adapted to gradually reduce the influence of human instruction. Such strategy was compared to a value-fixed setting, and Table 6 shows the comparison in one-round experiment.

The self-adapting strategy successfully generated 490 dialogues, where fixing K as 5 and 15 generated only 10 and 30 dialogues, respectively. When K is too small (K=5), the number of produced representative samples is limited, preventing the instruction pool from being updated sufficiently and diversely. This may quickly lead to stagnation, where the generation of corpus meeting the filtering criteria can no longer continue. When K is too large (K=15), the clustering may become overly dispersed, resulting in representative samples of inconsistent quality or excessive similarity, which similarly hampers the healthy evolution of the instruction pool and leads to inefficient generation. Numerically, the diversity metrics (Dt_1, Dt_2, *DD*, *SD*) under the fixed K-value settings are higher. However, this is due to the distortion of metrics caused by the excessively small sample sizes (10 and 30 dialogues). In an extremely small dataset, vocabulary and themes have little opportunity to repeat, artificially inflating the ratio of "non-repeating items/total items." Therefore, these inflated diversity values do not prove the superiority of the fixed K-value strategy. Instead, they indirectly confirm that such settings lead to the generation of unreliable small-scale datasets.

We also compared the effect of different g_t values in one-round experiment, as reported in Table 7. Setting g_t 2 achieves the optimal balance between scale and quality. It generated 490 dialogues and achieved the best overall performance across various diversity metrics. Setting g_t 1 also generated a substantial amount of data (374 dialogues), but its diversity metrics (particularly *DD* and *SD*) were slightly lower. Setting g_t 4 led to severe over-filtering, ultimately retaining only 95 dialogues. Its exceptionally high diversity metric values are statistical artifacts caused by the small sample size. This demonstrates that overly strict filtering criteria significantly impair the efficiency and scale of data generation. Therefore, in order to determine the optimal threshold, the following aspects should be considered: ensuring the basic interactive structure of dialogues, effectively filtering out low-quality samples, and maximizing the scale and efficiency of data generation.

In addition, we have observed that most of the failure cases in the experiment were filtered by the dialogue length threshold. Without the limitation or filtering by dialogue length, the generated dialogue may become too long and degrade significantly as LLM can fall into a meaningless repetitive cycle in the later stages of the dialogue, resulting in lengthy and ineffective text content, e.g., the doctor keeps saying similar contents such as "Wish you all the best" or repeats asking similar questions. Overall, the post-processing workflow successfully captures and removes such samples through the parameter thresholds, ensuring the quality of the final dataset.

Conclusions

This paper presents an automated framework to generate medical dialogue corpus for Medical Chinese Test. The framework involves building manual exemplar instructions and an automated pipeline for iteratively refining machine-generated instructions. To ensure diversity and relevance in the instructions, we introduce Maximal Marginal Relevance (MMR) and K-means clustering for effective sampling. By incorporating domain-specific knowledge and adopting the sampling strategy, we have successfully constructed a high-quality medical dialogue dataset of 20k examples that basically meets the guidelines of Medical Chinese Test (MCT). In terms of both objective and subjective indicators, the generated dataset MCT-Chat is proved comparable to real-world medical datasets.

In the current framework, each of the main four stages can be further refined towards better performance. For example, K-means can be replaced with more sophisticated clustering approaches, e.g. graph clustering for its ability in capturing complex semantic dependencies. In the future, we would like to expand the framework to multilingual and multicultural medical dialogues, ensuring alignment with regional guidelines while preserving the data quality. Besides, it is worth considering the integration of dynamic knowledge updates to adapt to evolving scenarios, and such enhancements can further improve the applicability of the generated corpora in real-world medical settings.

Data availability

The dataset sample (in English) generated during the current study are available via https://github.com/haozhou2018/dataset-mctchat. The original complete dataset (in Chinese) are available from the corresponding author on reasonable request.

Appendix

Here provide a few examples of manual instructions that emphasize certain components from the instruction subject description set defined in Section 4.2.

- Simulate a doctor-patient consultation dialogue, in which the doctor needs to understand the patient's medical history through questions, such as the location of the pain, the intensity of the pain, how long it has been present, etc. The number of dialogue turns is limited to 5 turns. (Component 6 and 8)
- Generate a discussion about disease diagnosis based on the scenario of doctor-doctor communication. In it, two doctors talk about a complex case in the emergency room, discussing possible diagnoses. The number of dialogue turns is 8. (Component 6, 7, and 8)
- Create a dialogue for a doctor-patient diagnosis, where the doctor tells the patient their diagnosis and explains the possible causes of the disease. The number of dialogue turns is 4. The instruction contains 20 words. (Component 6, 8, and 9)
- Generate a dialogue for a doctor-nurse communication, in which a doctor is explaining the steps of a clinical operation to nurse. The number of dialogue turns is limited to 6 turns. (Component 6 and 8)
- Set up a scenario where a nurse explains the details of a treatment plan and possible side effects to a patient during a patient-nurse consultation. Developed in the form of questions and answers. (Component 3, 6, and 7)
- Simulate a doctor-patient communication in which the doctor needs to tell the patient about their treatment plan, including the next steps in treatment and expected recovery time. The instruction length is within 20 words. (Component 6 and 9)
- Simulate a scene of a doctor-patient complaint, patients describes their main symptoms, and the doctor provides possible disease causes and further examination suggestions. The number of dialogue rounds is 5 turns. (Component 6, 7, and 8)
- You are a doctor, please inform the patient of the treatment plan for the disease in Chinese, the topic comes from the formulation of the diagnosis and treatment plan in the MCT task guidelines, and the dialogue is limited to 8 turns. (Component 5, 6, and 8)
- Please generate a dialogue about questions and answers between doctors and patients. The topic of the dialogue is clinical operations. The dialogue should include two turns. In the first turn, the doctor asks the patient whether he is willing to undergo surgery, and in the second turn, the doctor tells the patient the risks and benefits of surgery. (Component 6, 8, and 10)
- You are a nurse, please inform the patient about the disease prevention method in written language, the topic is the disease prevention in the MCT task guidelines, and the dialogue is limited to 4 turns. (Component 3, 6, and 8)

Received: 6 August 2025; Accepted: 31 December 2025
Published online: 06 March 2026

References

1. Zeng, G. *et al.* MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 9241–9250 (Online, 2020).
2. Joshi, A. *et al.* Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 3755–3763 (2020).
3. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.***30**, 2613–2622 (2024).
4. Tian, D. et al. The role of large language models in medical image processing: A narrative review. *Quant. Imaging Med. Surg.***14**, 1108–1121 (2024).
5. Yang, X. et al. A large language model for electronic health records. *NPJ Digital Med.***5**, 194 (2022).
6. Lucas, H. C., Upperman, J. S. & Robinson, J. R. A systematic review of large language models and their implications in medical education. *Med. Educ.***58**, 1276–1285 (2024).
7. Kumichev, G. *et al.* Medsyn: Llm-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 215–230 (Springer, 2024).
8. Zhong, Y. Advancing medical multimodal learning and data generation with diffusion model and llm. *Proc. AAAI Conf. Artif. Intell.***39**, 29319–29320 (2025).
9. Lu, K. *et al.* #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following* (2023).
10. Bach, S. *et al.* Promptsource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 93–104 (Dublin, Ireland, 2022).
11. Wang, Y. *et al.* Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5085–5109 (Abu Dhabi, United Arab Emirates, 2022).
12. Honovich, O. *et al.* Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 14409–14428 (Toronto, Canada, 2023).
13. Wang, Y. *et al.* Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13484–13508 (Toronto, Canada, 2023).
14. Sun, Z. *et al.* Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Advances in Neural Information Processing Systems*, 2511–2565 (2023).

15. Li, D. *et al.* Llm-ddi: Leveraging large language models for drug-drug interaction prediction on biomedical knowledge graph. *IEEE Journal of Biomedical and Health Informatics (Early Access)* .
16. Zheng, Z., Wang, L., Qiao, S., Zhou, Z. & Chen, J. Intelligent medical consultation system based on the gpt model. In *Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering*, 1192–1197 (2023).
17. Fawzi, S. A review of the role of chatgpt for clinical decision support systems. In *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 439–442 (IEEE, 2023).
18. Karak, A. & Kunal, K. Implementation of gpt models for text generation in healthcare domain. In *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, 1, 1–6 (IEEE, 2023).
19. Li, Y. *et al.* Chatdoctor: A medical chat model fine-tuned on LLaMA model using medical domain knowledge (2023). arXiv:2303.14070.
20. Das, T., Albassam, D. & Sun, J. Synthetic patient-physician dialogue generation from clinical notes using llm (2024). arXiv:2408.06285.
21. Liu, H. *et al.* LogiCoT: Logical chain-of-thought instruction-tuning data collection with GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2908–2921 (2023).
22. Liu, W. *et al.* Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 447–459 (Springer, 2022).
23. Bao, Z. *et al.* Disc-medllm: Bridging general large language models and real-world medical consultation (2023). arXiv:2308.14346.

Acknowledgements

Xiamen Natural Science Foundation (No. 3502Z202471035); Huaqiao University's Academic Project Supported by the Fundamental Research Funds for the Central Universities (No. 2024HQYJ01).

Author contributions

Hao Zhou conceptualized the idea and wrote the main manuscript. Xuanlang Hu made equal contribution to the first author, refining the framework and co-writing the manuscript. Ting He supervised the project. Haibin You and Jiyu Yin conducted the main experiment and made figures. Jingjing Xu and Zhizhong Lin conducted the ablation experiment. Huazhen Wang verified the experiment results and provided the basic analysis.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.Z. or T.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026