# Benchmarking Large Language Models on Communicative Medical Coaching: A Dataset and a Novel System

**Hengguan Huang**[1]*, **Songtao Wang**[1]*, **Hongfu Liu**[1], **Hao Wang**[2], **Ye Wang**[1]

[1]National University of Singapore, [2]Rutgers University

## Abstract

Traditional applications of natural language processing (NLP) in healthcare have predominantly focused on patient-centered services, enhancing patient interactions and care delivery, such as through medical dialogue systems. However, the potential of NLP to benefit inexperienced doctors, particularly in areas such as communicative medical coaching, remains largely unexplored. We introduce "ChatCoach," a human-AI cooperative framework designed to assist medical learners in practicing their communication skills during patient consultations. ChatCoach [1] differentiates itself from conventional dialogue systems by offering a simulated environment where medical learners can practice dialogues with a patient agent, while a coach agent provides immediate, structured feedback. This is facilitated by our proposed Generalized Chain-of-Thought (GCoT) approach, which fosters the generation of structured feedback and enhances the utilization of external knowledge sources. Additionally, we have developed a dataset specifically for evaluating Large Language Models (LLMs) within the ChatCoach framework on communicative medical coaching tasks. Our empirical results validate the effectiveness of ChatCoach.

## 1 Introduction

The advent of Natural Language Processing (NLP) has significantly impacted the healthcare domain, carving pathways for numerous applications that enhance both patient-centered services and healthcare operations. These applications encompass medical dialogue systems, automated medical coding, clinical decision support, and information extraction from electronic health records, among others (He et al., 2023). Despite the strides made in these areas, there remains largely untapped potential of NLP in aiding the professional development of early-stage medical learners and early-career practitioners. A critical aspect of this professional development revolves around enhancing communication skills, especially in the context of medical consultations.

A wealth of research highlights the critical importance of effective communication in medical practice. Choudhary and Gupta (2015) found a strong consensus among medical students on the need to refine communication skills for better medical practice, with a significant proportion showing marked improvements after training. Various other studies (Choudhary and Gupta, 2015; Chi and Wylie, 2014; Ruiz et al., 2006; Sargeant et al., 2010) have consistently shown that proficient communication skills are key to increasing patient satisfaction, enhancing diagnostic precision, and fostering stronger doctor-patient relationships. Despite this acknowledgment, the area of communicative medical coaching, particularly through leveraging advanced Language Language Models (LLMs), remains relatively unexplored.

Addressing this gap, we introduce *ChatCoach*, a novel human-AI cooperative framework devised to enhance communicative proficiency among medical learners. Unlike traditional dialogue systems focused on *patient* engagement, ChatCoach transitions the focus towards the professional development of *medical practitioners*. This approach fosters a dynamic environment where learners can engage in realistic dialogues, receive immediate feedback, and refine their understanding of medical terminologies. ChatCoach provides a simulated, realistic environment for medical learners to practice their communication skills during patient consultations. The architecture of ChatCoach (shown in Fig. 1(a)) includes a patient agent simulating real-world doctor-patient interactions and a coach agent providing real-time feedback on learners' termino-

* Equal contribution. Correspondence to: Hengguan Huang <huang.hengguan@u.nus.edu>

[1]Our data and code are available online: https://github.com/zerowst/Chatcoach



*Findings of the Association for Computational Linguistics ACL 2024*, pages 1624–1637
August 11-16, 2024 

logical usage.

A major challenge in this direction is the absence of publicly available data for communicative medical coaching, largely due to the sensitive nature of healthcare information and the substantial costs of data collection and annotation. To overcome this challenge, we devised a multi-agent data generation framework (shown in Fig. 1(b)) using external resources to produce training data for fine-tuning an open-source LLM. This framework employs LLM-based agents, including patient, coach, and doctor agents, which interact by querying and retrieving information from two sources: a medical dialogue database and a medical knowledge database. Additionally, we compiled a human-annotated testing dataset to assess LLMs' capabilities in communicative medical coaching.

Our contributions are threefold:

- We pioneer the utilization of LLMs for communicative coaching in healthcare, forging a novel intersection among education, healthcare, and AI.
- We introduce the first benchmark dataset and evaluation metrics for communicative medical coaching, enabling the assessment of LLMs coaching efficacy in a simulated practice environment.
- We present a new prompting strategy, dubbed as Generalized Chain-of-Thought (GCoT), devised to improve the generation of structured feedback and the incorporation of external knowledge, without the need for manually constructing reasoning steps. Our GCoT method demonstrates superior performance over various existing Chain-of-Thought techniques across tasks within our dataset.

## 2 Related Work

### 2.1 Medical NLP Applications with LLM

The field of healthcare has seen notable changes in recent years, driven in part by advances in Natural Language Processing (NLP) technologies. Initially, research efforts were concentrated on fundamental tasks such as Named Entity Recognition (NER) (Zhang et al., 2021; Nesterov and Umerenkov, 2022), Relation Extraction (RE) (Deng et al., 2020; Zhao et al., 2022), and Electronic Health Records (EHR) (Yu et al., 2019). These tasks posed challenges due to limited data access and the intricate nature of the medical domain. However, with the emergence of large language models (LLMs), the focus has shifted towards more practical applications, including the development of medical dialogue systems (Dou et al., 2023; Qin et al., 2023), innovative medical consultation platforms (Shi et al., 2023), and automated generation of medical reports (Zhao et al., 2023). Despite the strides made, the majority of existing models and tools primarily cater to patient-centered services. Notably absent are resources tailored for inexperienced medical learners and early-career doctors, a gap that our research seeks to address. This work delves into the potential of LLMs in enhancing the communication skills of medical professionals.

### 2.2 Medical Education with NLP

Traditional techniques aimed at enhancing communication skills include computer-assisted language learning (Levy, 1997), pronunciation training (Li et al., 2016), and mispronunciation localization (Wei et al., 2022b). These approaches typically rely on advanced acoustic models (Mohamed et al., 2011; Huang et al., 2019, 2020, 2021) to identify pronunciation errors and generate feedback. However, these applications are generally designed for the broader public and may not be ideally suited for clinical environments.

In a different vein of research, studies such as (Denny et al., 2003; Da Silva and Dennick, 2010; Zhang et al., 2012; Chary et al., 2019) have employed NLP techniques to enhance medical education by focusing on content analysis and student performance evaluation. Unlike these approaches, the current work introduces real-time coaching in communication skills specifically tailored for medical consultations. Utilizing Large Language Models (LLMs), it offers immediate, structured feedback, distinguishing itself from the predominantly static and retrospective analyses found in previous work.

### 2.3 Prompting-based Method

Prompting-based methods in Large Language Models (LLMs) have emerged as a versatile mechanism to guide models towards task-specific responses. Among the various strategies, in-context learning (Brown et al., 2020), where relevant examples are provided to tailor the model's behavior, and instruction prompting (Wang et al., 2022; Ouyang et al., 2022), where explicit task instructions are embedded within the prompts, have gained prominence. A notable advancement in this domain is the Chain-of-

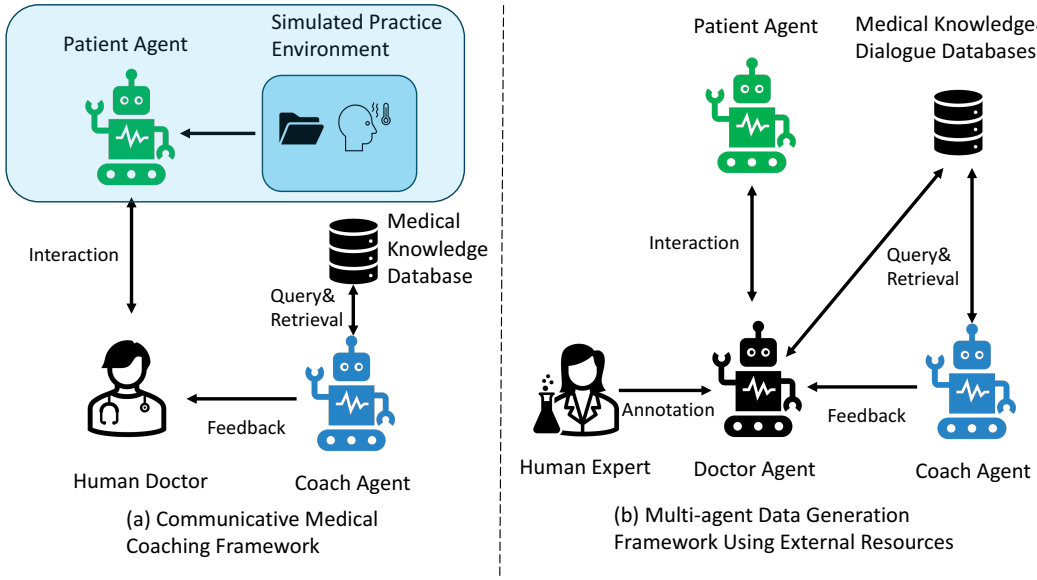


Figure 1: (a) General framework of communicative medical coaching. (b) Multi-agent data generation framework using external resources.

Thought (CoT) paradigm (Wei et al., 2022a), which introduced a chain of reasoning steps for each exemplar of in-context learning, significantly enhancing performance on complex reasoning tasks. Despite its advancements, CoT's reliance on human-crafted reasoning paths limits its applicability in open-ended settings, such as ours.

Following this, a variety of strategies have been proposed to improve upon the CoT paradigm. For instance, the zero-shot CoT (Kojima et al., 2022) extends the CoT paradigm to handle tasks by simply adding "think step by step" to the prompt, without requiring any exemplars and reasoning steps. However, such a method does not adequately integrate external knowledge or produce structured feedback that professionals can easily interpret, as observed in our human evaluation. In contrast, our GCoT introduces generalizable variables into the reasoning paths, enabling the generation of structured feedback and the effective integration of external knowledge. Additionally, the development of Auto-CoT (Zhang et al., 2022; Shum et al., 2023) aims to lessen the manual burden associated with formulating reasoning steps. However, this method's reliance on generating multiple samples from LLMs introduces computational inefficiencies and falls short in scenarios necessitating immediate feedback, such as our problem settings. This highlights the pressing need for solutions like GCoT that cater to real-time application requirements while enhancing the integration of external knowledge sources.

## 3 Communicative Medical Coaching

### 3.1 Problem Formulation

Given a medical knowledge database $\mathcal{D}$, which consists of a set of diseases $D = \{d_k \mid k = 1, \ldots, K\}$, where each $d_k$ includes a comprehensive description of the disease involving symptoms, medications, and other relevant clinical information. We define the simulated medical environment as $\mathcal{E}$, comprising a collection of scenarios $\mathcal{E} = \{e_j \mid j = 1, \ldots, J\}$. Each scenario $e_j = \{p_j, D_j\}$ corresponds to a *patient agent*, which encapsulates a patient profile $p_j$ and a specific medical context drawn from a subset of diseases $D_j \subseteq D$. The goal is to construct a simulated practice environment where a human doctor (i.e., a medical learner) can engage in medical dialogue with a *patient agent*. Concurrently, a *coach agent* delivers real-time feedback to the doctor.

### 3.2 System Overview

Figure 1(a) shows the architecture of the proposed system. It consists of two primary components: a patient agent, and a coach agent. The patient agent and the coach agent are driven by LLMs. The human doctor interacts with the patient agent in a simulated medical environment that is specified by each unique scenario $e_j$.

The patient agent generates responses $\mathcal{R}_j$ during the consultation based on the patient profile $p_j$, the current input from the doctor $S_j$, and the preceding

part of the conversation $\mathcal{H}_j^-$:

$$g(S_j, p_j, \mathcal{H}_j^-) \rightarrow \mathcal{R}_j, \quad (1)$$

where $\mathcal{H}_j^-$ represents the historical dialogue excluding interactions from the coach agent, ensuring that the coach agent's contributions do not affect the patient agent's responses.

Simultaneously, the coach agent monitors the dialogue between the human doctor and the patient agent, ready to provide feedback. This feedback mechanism is written as:

$$f(S_j, e_j, \mathcal{H}_j) \rightarrow \mathcal{F}_j, \quad (2)$$

where $f$ processes the doctor's dialogue $S_j$ and the complete history of the conversation $\mathcal{H}_j$ (including the coach agent's feedback in the previous dialogue round) within the context of $e_j$ to generate feedback $\mathcal{F}_j$. The purpose of the coach agent's feedback is to foster improved communication strategies by the doctor, such as correcting errors in medical terminology and providing constructive guidance and encouragement for more effective patient interactions. Both $g$ and $f$ are implemented by prompting LLMs.

## 4 Generalized Chain-of-Thought (GCoT)

Communicative medical coaching poses a unique challenge for Large Language Models (LLMs), characterized by its open-ended, knowledge-intensive reasoning demands. The feedback generated must be real-time and easily understandable by medical practitioners. Additionally, the reasoning process requires the utilization of external knowledge databases. Traditional prompting methods, such as zero-shot CoT often fall short in generating structured feedback and effectively incorporating external knowledge (refer to Fig. 3 for an example of coach feedback generated by prompting-based approaches). Here, we introduce the Generalized Chain-of-Thought (GCoT) approach. GCoT improves upon CoT by embedding generalizable variables within reasoning paths. These variables are elements shared across various data samples' reasoning steps, facilitating the creation of structured feedback and seamless external knowledge integration.

GCoT adopts a two-step process aimed at utilizing generalizable variables for prompt generation:

1. **Inferring Generalizable Variables across Data Samples:** The process begins with extracting generalizable variables from various input-output samples. This is accomplished by prompting an LLM with: "Imagine you are reasoning step by step from input to output, please infer generalizable variables in the reasoning steps across the following data samples." The input includes the doctor's statement and medical context from a medical knowledge database, with the output being the coach's feedback. This step is critical for identifying variables that represent both the conversation structure and the external knowledge sources, as depicted in Table 1.

2. **Prompt Generation Based on Inferred Variables:** After identifying these variables, the next step involves generating tailored prompts. The LLM is instructed with: "Generate the corresponding prompt for GPT-3.5, which should: (1) follow the Chain-of-Thought patterns; (2) ensure reasoning steps are not specific to any data; (3) base reasoning steps on these variables." This ensures the feedback (1) adheres to CoT step-by-step reasoning pattern, (2) abstracts reasoning steps for various data samples, and (3) incorporates the identified variables. The outcomes of this prompt generation are documented in Table 2, with variables indicated in square brackets, for instance, [incorrect symptom].

## 5 Constructing the ChatCoach Dataset: A Multi-Agent Approach for Generating Domain-Specific Conversational Data

The development of novel NLP applications, especially in specialized fields such as medical coaching, is hindered by the scarcity of domain-specific conversational datasets. To bridge this gap, we introduce a novel multi-agent data generation framework (illustrated in Fig. 1(b)) leveraging external resources to synthetically produce training data for fine-tuning open-source LLMs.

Our framework integrates three types of LLM-based agents – Patient, Coach, and Doctor – each designed to simulate real-world roles within medical dialogues. These agents interact through querying and retrieving information from two primary sources: a medical dialogue database and a medical knowledge database. Specifically, the Patient Agent simulates patient inquiries; the Doctor Agent generates medical responses, potentially including common errors; and the Coach Agent offers correc-

**Generalizable Variables Inferred by GCoT**

**Generalizable Variables across Examples:**

*Condition Miscommunication:*

- Incorrect disease name or symptom.
- Correct disease name or symptom based on medical context.

*Medication Miscommunication:*

- Incorrect medication name or treatment suggestion.
- Correct medication name or treatment suggestion based on medical context.

*Treatment Miscommunication:*

- Incorrect treatment advice.
- Correct treatment advice based on medical context.

Table 1: Generalizable variables inferred by GCoT.

**Generalized Chain-of-thought (GCoT)**

**Instruction**: As a linguistic coach for a junior doctor, your task is to evaluate the doctor's statement: {doctor's statement} against the provided medical context: {Medical Context}. Your evaluation should identify any discrepancies within the doctor's communication. Where discrepancies arise, guide the doctor towards more accurate medical terminology and understanding. If the statements align well with the medical context, provide positive reinforcement and additional advice if necessary.

**Thinking steps:**

*Identify Key Medical Terms:*

Extract medical terms from the doctor's statement, including diseases, symptoms, medications, and treatments.

*Compare with Medical Context:*

Check these terms against the medical context for accuracy in:

- Disease/symptom identification.
- Medication/treatment recommendation.

*Feedback:*

- *If Incorrect:* Point out the error and provide the correct term from the medical context. Use simple corrections like "Instead of [incorrect symptom], it should be [correct symptom]", "Instead of [incorrect medication name], it should be [correct medication name]" or "Instead of [incorrect disease name], it should be [correct disease name]".
- *If Correct:* Affirm with "Your diagnosis/treatment aligns well with the medical context. Good job."

Note: <correct symptom>, <correct medication name> and <correct disease name> are extracted from medical context

Table 2: GCoT prompt for ChatCoach.

tive feedback or encouragement, drawing from the medical knowledge database.

More importantly, to rigorlously evaluate the LLM's performance in medical coaching, we compiled a human-annotated testing dataset based on the aforementioned data.

**Data Generation Conditioned on External Resources** Recent studies, such as Jentzsch and Kersting (2023), have identified limitations in current LLMs' ability to generate diverse and contextually rich data samples. Our methodology addresses these limitations by conditioning the data generation process on external resources: the Disease database(Yunxiang et al., 2023), which encompasses comprehensive disease-related information (e.g., symptoms, diagnostic tests, treatments, and medications), and the MedDialog database(He et al., 2020), a corpus of real-world medical consultations. The inclusion of a coaching role, absent in MedDialog, and the simulation of doctor's

errors—uncommon in existing dialogues—pose unique challenges, which our framework overcomes by initiating data generation with patient queries from the MedDialog dataset. The Doctor Agent intentionally incorporates common misconceptions to simulate early-stage medical training errors. The Coach Agent then evaluates these responses against accurate medical statements, correcting terminological inaccuracies and enriching the dialogue with supportive insights for diagnostic reasoning.

### 5.1 Task Descriptions

The ChatCoach Dataset aims to benchmark LLMs' medical coaching efficacy, facilitating the development of communicative medical coaching tools for early-career doctors. We introduce two key tasks for assessing the quality of generated coaching feedback:

- **Detection of Medical Terminology Misuse:** This task involves identifying incorrect medical terminology in the doctor's responses, such as inappropriate disease diagnoses, irrelevant symptoms, or incorrect medication or test usage. Success depends on analyzing conversational history and applying relevant medical knowledge.
- **Correction of Medical Terminology Misuse:** Following the Detection Task, this task focuses on providing corrective advice to address any identified terminology misuse. It similarly requires a deep understanding of conversational context and medical knowledge.

Evaluating the coach's feedback for constructiveness, knowledgeability, and clarity is also crucial, although these aspects present quantification and evaluation challenges. We plan to explore these dimensions in future work.

**Human Annotation** For initial annotation, we engage 2-3 annotators (either medical professionals or knowledgeable students) to review doctor responses within 500 conversations, including patient, doctor, and coach interactions. Annotations focus on the detection and correction of medical terminology misuse, with coach feedback serving as a reference. To ensure quality, we manually validate each annotation, utilizing advanced LLMs like GPT-4 to calculate inter-rater agreement rates. We pay special attention to conversations with low agreement, assessing the plausibility, relevance, and completeness of annotations. Ultimately, from the initial 500 conversations, we retain 291 based on rigorous evaluation criteria.

| Statistics | Number |
|---|---|
| Total conversation | 291 |
| Disease | 99 |
| Doctor's statement | 1,315 |
| Patient's response | 1,315 |
| Condition | 166 |
| Disease | 98 |
| Medication | 39 |
| Treatment case | 295 |
| Correction case | 291 |
| Nonlingual case | 1,015 |

Table 3: Statistics of Testing set.

### 5.2 Dataset Overview

The ChatCoach dataset comprises 3,500 conversations with 13,666 utterances, based on real-world medical consultations from the MedDialog dataset. The dataset is divided into training (2,509 conversations), validation (700), and testing (291) sets. While training data may not be essential for benchmarking closed-source LLMs like ChatGPT, it is crucial for fine-tuning less capable models. The testing set covering 99 diseases, thoroughly annotated, serves as the primary resource for evaluating LLMs' medical coaching performance. Detailed statistics of this set are provided in Table 3, highlighting the distribution of conversations, doctor and patient statements, the numbers of condition, medication, and treatment miscommunication errors, and the categorization of cases into detection, correction, and non-linguistic advice. Notice that "Nonlingual cases" in Table 3. corresponds to non-linguistic advice expected from the Coach Agent when no direct medical terminology errors occur in the doctor's statement. In such scenarios, the coach might provide encouragement, further medical insights, or advice to progress the diagnostic procedure or maintain the dialogue's flow.

## 6 Experiments

We assess the medical coaching capabilities of both an open-sourced LLM and a close-sourced LLM using the proposed ChatCoach frameworks on two tasks specified in Sec. 5, namely detection and

| Method | Detection | | | Correction | | |
|---|---|---|---|---|---|---|
| | BLEU-2 | Rouge-L | BERTScore | BLEU-2 | Rouge-L | BERTScore |
| Training-based | | | | | | |
| Instruction-Tuning | **39.8** | 3.0 | **77.8** | **4.0** | 1.7 | 59.7 |
| Prompting-based | | | | | | |
| Instruction Prompting | 27.4 | 3.3 | 67.6 | 1.4 | 2.1 | 61.6 |
| Vanilla CoT | 17.7 | 2.7 | 64.1 | 0.1 | **2.3** | 58.1 |
| Zero-shot CoT | 27.6 | 1.9 | 69.0 | 3.0 | 0.9 | 58.8 |
| GCoT (Ours) | **34.2** | **3.7** | **72.4** | 1.6 | 2.0 | **65.4** |
| Human | 76.6 | 6.0 | 90.5 | 33.5 | 3.6 | 84.1 |

Table 4: Performance comparison of various methods on the detection and correction of medical terminology errors.

correction of misuse of the medical terminology. The generated coach feedbacks are evaluated based on both automatic and human evaluation metrics.

### 6.1 Experiment Setup

**Baselines** We investigate the following methods for addressing our problem settings:

- Vanilla Instruction Prompting: A method where the LLM is prompted with direct instructions for dialogue generation without further context.
- Zero-shot Chain of Thought (CoT) (Kojima et al., 2022): A simple CoT approach where the LLM is prompted with instructions for dialogue generation, being asked to generate a reasoning chain step by step.
- Vanilla Chain of Thought (Wei et al., 2022a): An extension of CoT where the model is given a few examples involving the corresponding reasoning path.
- Instruction Tuning (Longpre et al., 2023): A training-based method that includes instructions to the training input-output pairs for fine-tuning LLMs.

**Evaluation Metrics** For the quantitative study, since both detection and correction tasks belong to natural language generation, we employ conventional metrics, including BLEU-2, ROUGE-L, and BERTScore. BLEU-2 measures the precision of bi-gram overlaps, offering insights into the lexical accuracy of the generated text against reference answers. ROUGE-L assesses sentence-level similarity, focusing on the longest common subsequence to evaluate structural coherence. BERTScore is used for a semantic similarity assessment, utilizing BERT embeddings to compare the generated outputs and reference texts on a deeper semantic level.

The generated feedback from Coach Agents comprises open-ended natural language text. We adopt GPT-4 to extract the medical terminology errors and the corresponding corrections from the Coach Agents' feedback, then calculate the automated metrics based on the extracted information against human annotations. To further validate whether the automatic metrics-based on our annotated reference answers align with the the actual quality of model predictions, we conduct additional human evaluation.

**Implementation Details** We adopt 'gpt-3.5-turbo' for all our prompting-based methods. The prompts for all experiments are detailed in the Appendix. For instruction-tuning, we adopt QLORA (Longpre et al., 2023) to fine-tune a variant of Llama2, named Chinese_Alpaca2_LORA_13B, using 4*A40 GPUs for approximately 9 hours, with a batch size of 64, a learning rate of $2 \times 10^{-4}$, and a maximum of 1000 training steps. ( See Appendix for prompts of our baseline approaches.)

### 6.2 Results

We present the performance of various methods in Table 4, focusing on the detection and correction of medical terminology errors. The apparent gap between machine-generated results and human benchmarks in all evaluated metrics signals the inherent challenges within communicative medical coaching.

| Error Category | Zero-shot CoT | GCoT |
|---|---|---|
| Overly Divergent Advice | 7.14 | 0.79 |
| Excessive Coaching | 5.56 | 3.97 |
| Limited Medical Knowledge | 5.56 | 1.59 |
| Role Mismatch | 1.59 | 0.00 |

Table 5: Error rate (%) comparison between zero-shot CoT and GCoT (ours).

**Detection of Medical Terminology Misuse** In terms of the detection task, the results demonstrate GCoT's effectiveness in identifying medical terminology errors, with our method achieving competitive scores in BLEU-2 (34.2), Rouge-L (3.7), and BERTScore (72.4) metrics. Despite the Instruction-tuning method's higher scores in some metrics, GCoT's performance stands out among other prompting methods, indicating its effectiveness without the need for additional fine-tuning.

**Correction of Medical Terminology Misuse** In the correction task, although Instruction-Tuning continues to lead in performance with a BLEU-2 score of 4.0, the gap narrows, indicating the intrinsic challenge associated with generating contextually accurate corrections. When evaluating with the BERTScore, GCoT showcases its strength with a notable BERTScore of 65.4, surpassing both the prompting-based method and the training-based method. This discrepancy indicates that Instruction-Tuning, despite its effectiveness in generating responses that structurally follow given response patterns, may not fully capture the semantic nuances required for the diverse range of correct responses. This might be due to the method's tendency to overfit to the examples within the training dataset, limiting its ability to generalize to the varied corrections encountered in the test set.

**Human Evaluation** To validate the previously observed results, we conducted a human evaluation. We randomly selected 10% (126 instances) of Testing set for this purpose. Feedback generated by both Baseline Zero-shot CoT and our GCoT was reviewed by two participants, who were asked to rate each piece of feedback on a scale from 1 to 4, with respect to constructiveness, clarity, knowledgeability, and overall quality. Table 6 shows the average scores for each criterion. The results clearly indicate that our proposed approach, GCoT, significantly outperforms the baseline Zero-shot CoT, particularly in terms of clarity and constructiveness. This underscores GCoT's effectiveness in producing structured feedback that is easily understandable for users.

| Metric | CoT | GCoT (ours) |
|---|---|---|
| Constructiveness | 2.41 | **2.68** |
| Clarity | 2.15 | **3.10** |
| Knowledgeability | 2.35 | **2.39** |
| Overall | 2.21 | **2.52** |

Table 6: Human evaluation of coach feedback generated by GCoT and Zero-shot CoT.

**Error Analysis** To delve deeper into the sources of errors within Zero-shot CoT and GCoT implementations, we annotated all instances involved in the human evaluation, categorizing them into four distinct classes:

- **Overly Divergent Advice:** Feedback that is too wide-ranging, long, or off-topic, reducing its effectiveness.
- **Excessive Coaching:** Feedback inappropriately critiques suitable responses for lacking professional jargon.
- **Limited Medical Knowledge:** Errors due to insufficient use or understanding of the medical knowledge database.
- **Role Mismatch:** Instances where feedback shifts from a coach's to a doctor's perspective, misaligning with the intended advisory role.

As demonstrated in Table 5, the comparison between Zero-shot CoT and GCoT reveals significant improvements across all error categories with our GCoT approach. Notably, GCoT dramatically reduces the incidence of overly divergent advice from 7.14% to a mere 0.79%. Similarly, errors categorized under Limited Medical Knowledge dropped from 5.56% to 1.59%. These results underscore GCoT's capability to generate more targeted and organized feedback while effectively utilizing external medical knowledge.

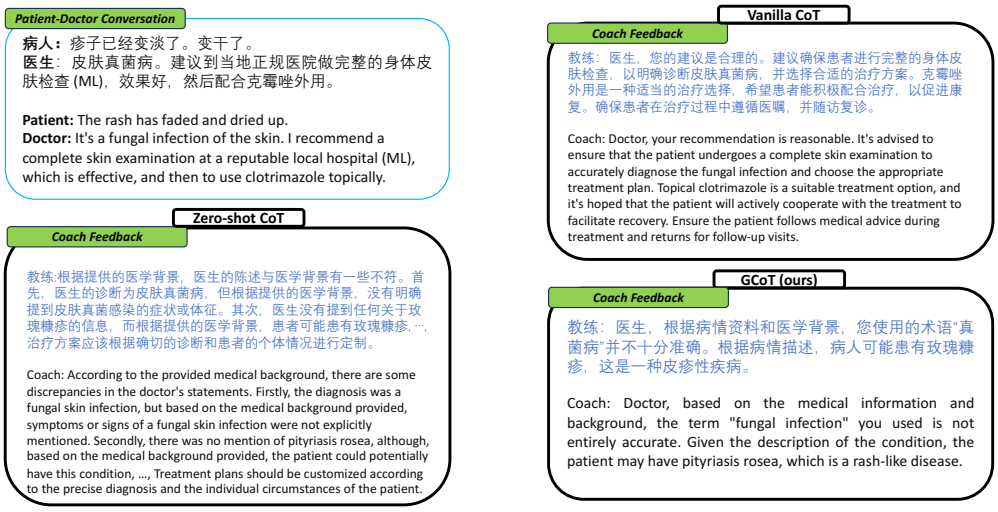


Figure 2: Example of coach feedback generated by various approaches. Vanilla CoT fails to identify errors in medical terminology, possibly due to lacking integration with external knowledge. While thorough, Zero-shot CoT generates overly verbose feedback unsuited for real-time application. In contrast, GCoT identifies errors effectively and provides concise and well-structured feedback, demonstrating superior integration of external medical knowledge for practical real-time coaching.

**Case Study** Figure 3 showcases an example of coach feedback generated by both baseline prompting methods and our Generalized Chain-of-Thought (GCoT) approach for comparison. In this case, we observe that Vanilla CoT is unable to detect errors in medical terminology, possibly due to its inadequate utilization of external medical knowledge. Zero-shot CoT, on the other hand, produces feedback that is lengthy and circuitous, making it less suitable for the immediacy required in real-time coaching environments. In stark contrast, the example illustrates how GCoT provides feedback that is notably more organized and precise, demonstrating its enhanced ability to integrate external medical knowledge sources effectively.

## 7 Conclusion

This work introduces ChatCoach, a new human-AI cooperative framework for communicative medical coaching. At the core of our approach is the Generalized Chain-of-Thought (GCoT), a strategy that significantly improves feedback structuring and the integration of external knowledge. We developed the first benchmark dataset designed to evaluate the medical coaching capabilities of Large Language Models (LLMs) within the ChatCoach framework. Through a series of automatic and human evaluations, we demonstrate ChatCoach's effectiveness in tackling two key tasks in communicative medical coaching, showcasing its potential to enhance medical education through AI.

## Limitations

Despite the advancements made by ChatCoach and the Generalized Chain-of-Thought (GCoT) approach, limitations persist that require further exploration and enhancement. Specifically, our error analysis reveals that GCoT still faces challenges with excessive coaching, where the system may critique acceptable responses for not using professional jargon. This indicates a need for refinement in distinguishing between instances that genuinely require correction. Addressing this issue involves developing an additional component within GCoT that accurately identifies when coach intervention is necessary, thereby reducing unwarranted critiques and enhancing the relevance and precision of the feedback provided.

## Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions. This project is funded by a research grant MOE-MOESOL2021-0005 from the Ministry of Education in Singapore.

## References


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Michael Chary, Saumil Parikh, Alex F Manini, Edward W Boyer, and Michael Radeos. 2019. A review of natural language processing in medical education. *Western Journal of Emergency Medicine*, 20(1):78.

Michelene TH Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243.

Anjali Choudhary and Vineeta Gupta. 2015. Teaching communications skills to medical students: Introducing the fine art of medical practice. *International Journal of Applied and Basic Medical Research*, 5(Suppl 1):S41.

Ana L Da Silva and Reg Dennick. 2010. Corpus analysis of problem-based learning transcripts: an exploratory study. *Medical education*, 44(3):280–288.

Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. 2020. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics*, 36(15):4316–4322.

Joshua C Denny, Jeffrey D Smithers, Randolph A Miller, and Anderson Spickard III. 2003. "understanding" medical school curriculum content using knowledgemap. *Journal of the American Medical Informatics Association*, 10(4):351–362.

Chengfeng Dou, Zhi Jin, Wenping Jiao, Haiyan Zhao, Zhenwei Tao, and Yongqiang Zhao. 2023. Plug-and-play medical dialogue system. *arXiv preprint arXiv:2305.11508*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, et al. 2020. Meddialog: Two large-scale medical dialogue datasets. *arXiv preprint arXiv:2004.03329*.

Hengguan Huang, Hongfu Liu, Hao Wang, Chang Xiao, and Ye Wang. 2021. Strode: Stochastic boundary ordinary differential equation. In *International Conference on Machine Learning*, pages 4435–4445. PMLR.

Hengguan Huang, Hao Wang, and Brian Mak. 2019. Recurrent poisson process unit for speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6538–6545.

Hengguan Huang, Fuzhao Xue, Hao Wang, and Ye Wang. 2020. Deep graph random process for relational-thinking-based speech recognition. In *International Conference on Machine Learning*, pages 4531–4541. PMLR.

Sophie Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Michael Levy. 1997. *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Kun Li, Xiaojun Qian, and Helen Meng. 2016. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):193–207.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. 2011. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22.

Alexander Nesterov and Dmitry Umerenkov. 2022. Distantly supervised end-to-end medical entity extraction from electronic health records with human-level quality. *arXiv preprint arXiv:2201.10463*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. *arXiv preprint arXiv:2305.05138*.

Jorge G Ruiz, Michael J Mintzer, and Rosanne M Leipzig. 2006. The impact of e-learning in medical education. *Academic medicine*, 81(3):207–212.

Joan Sargeant, Heather Armson, Ben Chesluk, Timothy Dornan, Kevin Eva, Eric Holmboe, Jocelyn Lockyer, Elaine Loney, Karen Mann, and Cees van der Vleuten. 2010. The processes and dimensions of informed self-assessment: a conceptual model. *Academic Medicine*, 85(7):1212–1220.

Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2023. Midmed: Towards mixed-type dialogues for medical consultation. *arXiv preprint arXiv:2306.02923*.

KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022a. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wei Wei, Hengguan Huang, Xiangming Gu, Hao Wang, and Ye Wang. 2022b. Unsupervised mismatch localization in cross-modal sequential data with application to mispronunciations localization. *Transactions on Machine Learning Research*.

Xin Yu, Wenshen Hu, Sha Lu, Xiaoyan Sun, and Zhenming Yuan. 2019. Biobert based named entity recognition in electronic medical record. In *2019 10th international conference on information technology in medicine and education (ITME)*, pages 49–52. IEEE.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Rui Zhang, Serguei Pakhomov, Sophia Gladding, Michael Aylward, Emily Borman-Shoap, and Genevieve B Melton. 2012. Automated assessment of medical training evaluation text. In *AMIA annual symposium proceedings*, volume 2012, page 1459. American Medical Informatics Association.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Bo-Wei Zhao, Lun Hu, Zhu-Hong You, Lei Wang, and Xiao-Rui Su. 2022. Hingrl: predicting drug–disease associations with graph representation learning on heterogeneous information networks. *Briefings in bioinformatics*, 23(1):bbab515.

Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. 2023. Chatcad+: Towards a universal and reliable interactive cad using llms. *arXiv preprint arXiv:2305.15964*.

## A A Failed Example of Coach Feedback

## B Prompts of Our Baseline Approaches

In this section, we present the prompts used for each baseline approach: Instruction Prompting (see Table 7), Vanilla Chain-of-Thought (see Table 8), and Zero-shot Chain-of-Thought (see Table 9).

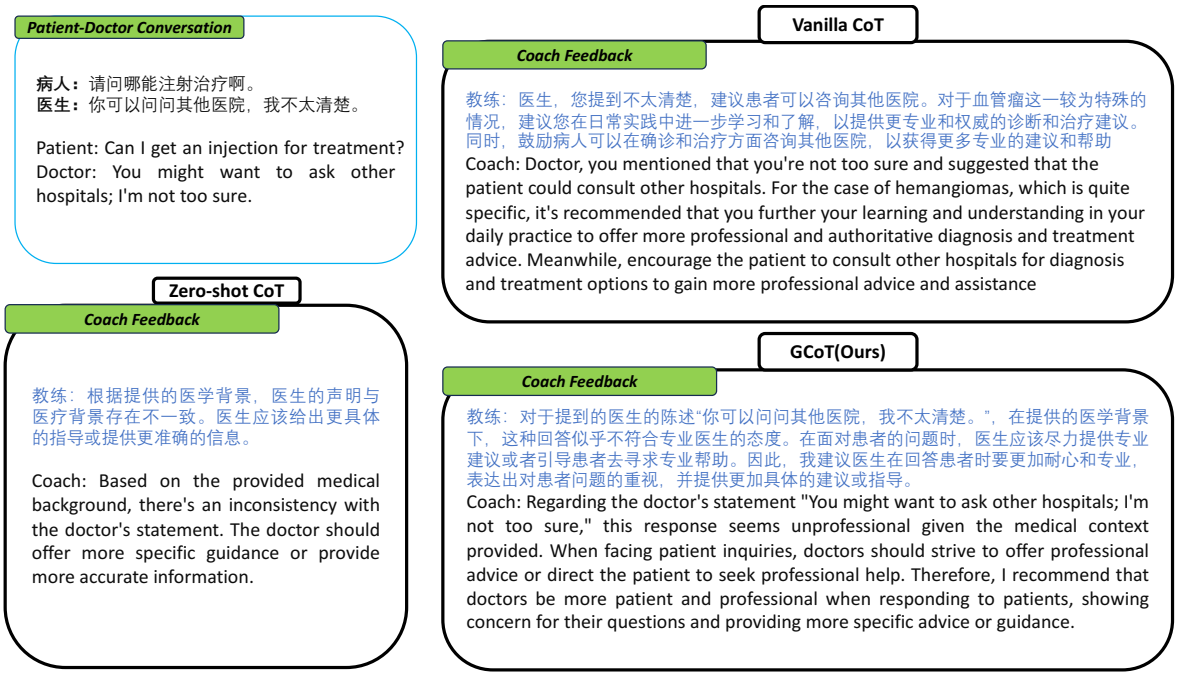


Figure 3: A failed example of coach feedback from various prompting-based approaches, demonstrating the issue of excessive coaching.

| **Vanilla Instruction Prompting** |
|---|
| **Instruction**: As a linguistic coach for a junior doctor, evaluate the doctor's statement: {doctor's statement} against the given medical context: {medical context}. If there are discrepancies, guide the doctor. If not, provide positive feedback. |

Table 7: Instruction prompting for ChatCoach.

| **Vanilla Chain-of-thought** |
|---|
| **Instruction**: As a linguistic coach for a junior doctor, evaluate the doctor's statement: {doctor's statement} against the given medical context: {medical context}. You should provide your response based on the following examples of input, thinking steps and output.<br>**Example 1**:<br>**Input**:<br>{doctor's statement for Example 1}<br>{medical context for Example 1}<br>**Thinking steps**:<br>{thinking steps for Example 1}<br>**Output**:<br>{coach's feedback for Example 1}<br>**Example 2**: {example2}<br>**Example 3**: {example3}<br>**Input**:<br>{doctor's statement}<br>{medical context} |

Table 8: Vanilla CoT for ChatCoach.

| **Zero-shot Chain-of-thought** |
| --- |
| **Instruction**: As a linguistic coach for a junior doctor, evaluate the doctor's statement: {doctor's statement} against the given medical context: {medical context}. If there are discrepancies, guide the doctor. If not, provide positive feedback.<br>**Please think step by step.** |

Table 9: Zero-shot CoT for ChatCoach