Advancing Chinese Conversation-based Patient Guidance with a Benchmark and Knowledge-Evolvable Assistant

Wenpeng Lu, *Member, IEEE*, Kangjun Liu, Jianlei Wang, Xueping Peng, *Senior Member, IEEE*, Tao Shen, Fa Zhu, Weiyu Zhang*, *Member, IEEE*, Jiabing Zhu, Tao Xin, Athanasios V. Vasilakos, *Senior Member, IEEE*

***Abstract*— Chinese Conversation-based Patient Guidance (CCPG) helps patients reach the correct hospital department through natural-language exchanges with medical staff. Despite the rapid success of large language models (LLMs) in other healthcare tasks, CCPG remains under-explored and lacks dedicated benchmarks. We address this gap with PG-Bench, the first comprehensive CCPG benchmark, spanning five subsets, 19,814 annotated dialogues, and 98 clinical departments. We evaluate 25 representative LLMs on PG-Bench and observe uniformly poor performance, even the latest models such as GPT-4 and DeepSeek-V3 fail to meet practical requirements. To close this gap, we introduce the Knowledge-Evolvable Assistant (KEA), a novel framework that augments any LLM with (i) an experience bank of validated, successful CCPG cases for analogy-based reasoning; (ii) a reflection bank that records previously misclassified cases together with their corrections and self-summarized error analyses; and (iii) an external medical knowledge base. KEA employs retrieval-augmented generation to evolve its guidance knowledge iteratively. Experiments show that KEA consistently and significantly boosts the CCPG performance of all tested LLMs on PG-Bench. However, current best results still fall short of clinical expectations, underscoring the difficulty of CCPG and the need for further research. PG-Bench and KEA together establish a rigorous foundation and strong baseline for future work on conversation-driven patient guidance in Chinese healthcare settings. Our code, datasets, supplementary experimental results, prompt templates, and department lists are available on our GitHub[1].**

***Index Terms*— Large Language Model, Healthcare Applications, Retrieval Augmented Generation, Patient Guidance**

Wenpeng Lu, Kangjun Liu, Jianlei Wang and Weiyu Zhang are with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), and are with Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Shandong Key Laboratory of Key Technologies and Systems for Humanoid Robots, Jinan, China (e-mail: wenpeng.lu@qlu.edu.cn, 10431240007@stu.qlu.edu.cn, 10431230002@stu.qlu.edu.cn, zwy@qlu.edu.cn).

Xueping Peng and Tao Shen are with the Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney (e-mail: xueping.peng@uts.edu.au, tao.shen@uts.edu.au).

Fa Zhu is with the College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing, China (E-mail: fazhu@njfu.edu.cn).

Jiabing Zhu is with Inspur Software Group Co., Ltd., Jinan, China (E-mail: zhujb@inspur.com)

Tao Xin is with the First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, Jinan, China (E-mail: xintao@sdfmu.edu.cn).

Athanasios V. Vasilakos is with the Department of ICT and Center for AI Research, University of Agder (UiA), Grimstad, Norway (E-mail: thanos.vasilakos@uia.no).

Weiyu Zhang is the corresponding author.

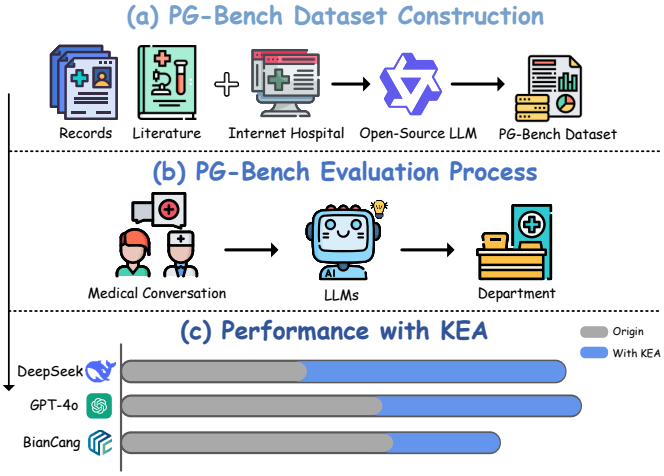

Fig. 1. PG-Bench evaluation suite, which establishes a novel benchmark for Chinese conversation-based patient guidance (CCPG), along with a new knowledge-evolvable assistant (KEA) framework.

I. Introduction

CHINESE conversation-based Patient Guidance (CCPG) plays a vital role in assisting patients in navigating to the appropriate departments by facilitating Chinese-language conversations between patients and hospital staff. Currently, Chinese hospitals typically assign front desk nurses to provide CCPG services, helping patients find the right department and receive timely treatment. However, the service heavily relies on the experience of nurses, requiring manual responses, which are time-consuming, labor-intensive, and burdensome. Moreover, as the number of patients increases, the increasing workload causes nurses to inevitably become fatigued, making

[1]https://github.com/KJOrigin/PG-Bench.

errors more likely. Clearly, there is an urgent need to implement an automated CCPG service. The development of medical artificial intelligence lays the foundation for addressing this issue and makes it possible [1].

In recent years, large language models (LLMs) have gained exceptional success in various domains, significantly transforming the landscape of medical artificial intelligence [2]–[5]. AI models have been widely applied across various aspects of healthcare, from assisting in clinical diagnosis to analyzing health records for personalized care, thereby significantly reducing the burden on medical professionals [6]–[18]. Although a great number of healthcare applications are being explored, conversation-based patient guidance has been neglected, despite its crucial role in helping patients navigate to the correct departments and receive timely treatment. In contrast, most existing works focus on triage, which prioritizes patients based on the severity of their conditions, rather than navigating them to the appropriate department [19]–[21]. With the aging population in China and the growing awareness of health management, the number of medical treatments in Chinese hospitals has surged rapidly. According to data from the National Health Commission of China, the total number of treatments reached 1.87 billion in just three months[2]. In the face of such immense healthcare demand, there is an urgent need for a CCPG system to provide effective support.

To fill this gap, in this work, we focus on the CCPG task, which aims to help patients navigate to the correct departments based on conversations in Chinese between patients and hospital staff. As shown in Figure 1, we first construct a novel CCPG benchmark (PG-Bench), which is built on dialogue generation templates and conversations from online hospitals. Then, we test the performance of various baselines on PG-Bench, including state-of-the-art general LLMs and domain-specific LLMs. Our findings indicate that existing methods struggle with this task. Finally, to address this challenge, inspired by *Agent hospital* [22], we propose a novel Knowledge-Evolvable Assistant (KEA) framework to adapt relevant Chinese medical guidance knowledge to LLMs with a self-evolution mechanism, without relying on parameter fine-tuning. Experimental results confirm that the KEA framework is highly compatible with any LLM and significantly enhances its performance, demonstrating promising capabilities. Our main contributions are summarized below:

- We propose a novel Chinese conversation-based Patient Guidance (CCPG) task. To the best of our knowledge, we are the first to investigate this task, which is essential to reducing hospital burden, improving the patient experience and enhancing healthcare efficiency.
- We introduce a novel CCPG benchmark (PG-Bench), which includes five subsets and contains a total of 19,814 instances across 98 departments. We test the performance of 25 widely used and state-of-the-art LLMs, none of which achieve satisfactory performance, highlighting the significant research potential of this task.
- We propose the Knowledge-Evolvable Assistant (KEA) framework, which elaborates a self-evolution mechanism to dynamically update knowledge and enhance the guidance ability of LLMs. KEA is compatible with any LLM, improving its performance without requiring training.

[2]https://www.nhc.gov.cn/mohwsbwstjxxzx/s7967/202410/3cf38fb0dc3043caba498867e744fa7e.shtml.

II. Related Work

A. Intelligent Triage

Intelligent triage, which prioritizes patients based on the urgency of their conditions, has been widely studied in healthcare [23]–[25], with approaches roughly categorized into two groups: (1) traditional machine learning-based methods and (2) deep learning-based methods.

Traditional machine learning methods, such as decision trees [26], support vector machines (SVMs), naive Bayes classifiers [27], and k-nearest neighbors [28], have been extensively applied to triage tasks. For example, SVMs [29] are used to classify patients into urgency levels based on symptom severity scores. However, these models fail to handle unstructured patient narratives or multi-turn dialogues due to rigid feature dependencies. To partially address this, hybrid frameworks combining rule-based systems with random forests [30] or gradient boosting machines (GBMs) [31] were proposed, but they still face challenges in understanding the diverse ways patients express their symptoms.

Deep learning approaches, including LSTMs [32], TextCNNs [33], and Transformer-based models [34], have advanced triage systems by automating feature extraction and improving semantic understanding. For instance, hierarchical LSTMs and GRUs were designed to model sequential dependencies in patient dialogues, while BERT-based models [35] fine-tuned on medical corpora, have achieved improved accuracy in symptom classification. TriageAgent [21] proposed a heterogeneous multi-agent framework that leverages LLMs to improve collaborative decision-making in clinical triage.

While traditional machine learning and deep learning methods have advanced the development of intelligent triage systems, and recent efforts have begun to explore automated patient guidance in clinical settings to enhance the medical experience, most studies focus mainly on prioritizing patients based on the severity of their conditions. In contrast, the exploration of conversation-based patient guidance has been overlooked, despite its critical role in helping patients navigate to the appropriate departments and receive timely care. Unlike conventional triage systems that prioritize patients according to clinical urgency, CCPG focuses on department-level navigation through natural-language conversations, aiming to optimize patient flow rather than emergency prioritization.

B. LLMs in Medical Domain

Recent advancements in medical large language models (LLMs) related to patient guidance have concentrated on three key applications: (1) disease diagnosis, (2) prognostic care, and (3) medical question answering.

Disease diagnosis models, such as BianCang [12], utilize pre-trained language models fine-tuned on clinical datasets to identify diseases from patient symptoms, histories, or diagnostic reports. These models achieve high accuracy in controlled settings by leveraging structured medical knowledge [36].

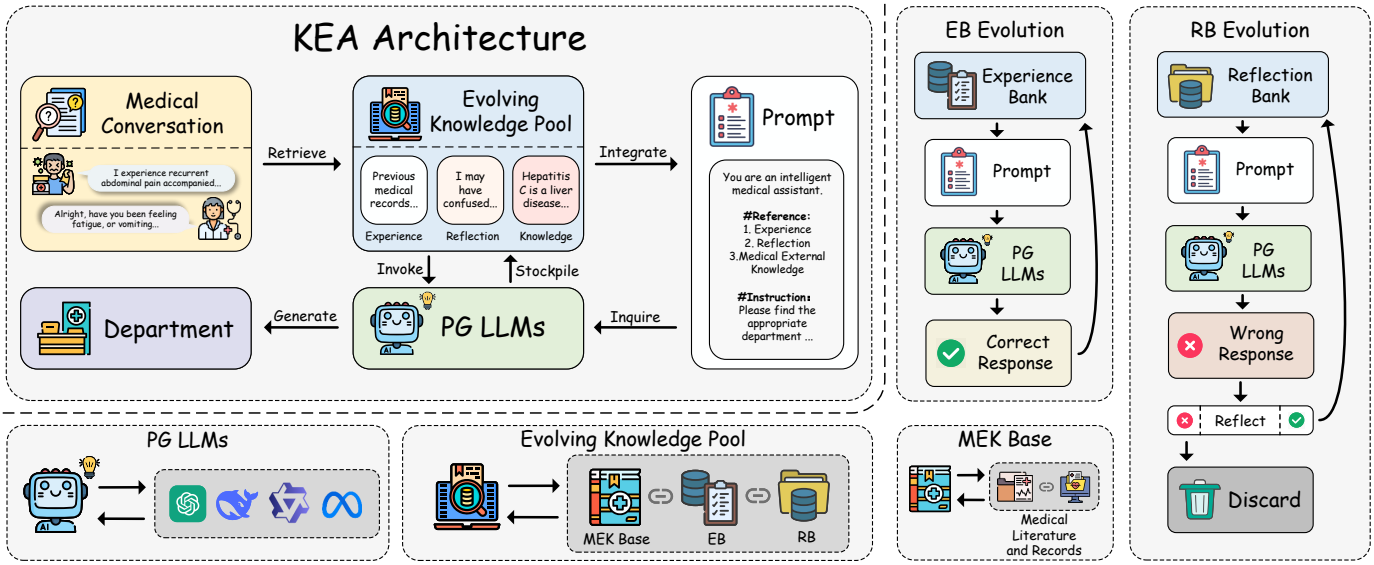

Fig. 2. Overall architecture of the proposed Knowledge-Evolvable Assistant (KEA) framework. KEA dynamically interacts with the medical guidance LLM through the evolving knowledge pool to enable CCPG. On the right side and the lower left corner of the figure are the components of KEA. EB: *Experience Bank*. RB: *Reflection Bank*. MEK Base: *Medical External Knowledge Base*. PG: *Patient Guidance*.

Prognostic care models, such as ChiMed-GPT [9], leverage patient data to predict treatment outcomes and recommend personalized care plans. By modeling the relationships between interventions and health trajectories, these systems aim to optimize clinical decision-making [37].

Medical question answering models, such as HuatuoGPT-o1 [10] and Sunsimiao, excel in answering exam-style questions and complex clinical queries. These models leverage large-scale biomedical corpora to generate relevant answers [38], [39]. Both HuatuoGPT-o1 and Sunsimiao have demonstrated exceptional performance on the CMB-Exam [40] benchmark.

Existing medical LLMs have been applied across various aspects of clinical practice. However, their performance on conversation-based patient guidance tasks remains unexplored. Moreover, most current medical LLMs rely on parameter fine-tuning to acquire domain-specific medical knowledge, which demands high-quality training data, substantial computational resources, and considerable human effort. Unlike conventional medical LLMs that rely on parameter fine-tuning for domain adaptation, KEA is a plug-and-play framework that can be seamlessly integrated with any LLM, enhancing its patient guidance capability through retrieval-augmented and self-evolving mechanisms without additional training.

III. Chinese Conversation-based Patient Guidance

A. Task Formulation

Given a medical conversation $D = \{u_1, \cdots, u_n\}$, where u_i represents an utterance from either a patient or hospital staff, the CCPG task is to infer the most appropriate departments $y = \{c_1, \cdots, c_m\}$ for the patient based on D, where $c_i \in C$, with C denoting the set of candidate departments in hospital.

B. CCPG Benchmark (PG-Bench)

As a novel task that has long been neglected, there are currently no available open-source datasets for the CCPG task. The privacy concerns of patients present a significant challenge in collecting real-world conversation datasets. To address the issue, we propose a template-based method (TBM) [41] that leverages LLMs to construct the benchmark (PG-Bench). Specifically, TBM consists of three stages: *Template design*, *Label alignment*, and *Conversation generation*.

Template Design: To simulate CCPG dialogues in real-world healthcare scenarios, we design four possible dialogue intents [41] : (1) description of the chief complaint from patients, (2) implicit symptom inquiry by hospital staff, (3) confirmation or denial of symptoms by patients, and (4) guidance to the target department by hospital staff. For each intent, multiple templates are designed to generate utterances. The template-based method enables systematic and controllable generation of diverse dialogues while preserving linguistic naturalness by leveraging real conversational patterns from hospitals.

Label Alignment: To ensure alignment between the patient's health condition and the appropriate department in the generated dialogues, we collect knowledge about the relations between symptoms, diseases, and departments from multiple online hospital databases[3], and construct symptom-disease-department triples based on these relations.

Conversation Generation: To generate complete and fluent dialogues, we first incorporate the symptom-disease-department triples into the dialogue templates — for example, ["cough, sputum production, dyspnea", "acute bronchitis", "Internal Medicine"] — and then leverage the large language model Qwen-max [42], accessed via a commercial API, to preliminarily construct the raw dialogue samples. Next, we invited two professional medical experts to evaluate the raw dialogue samples based on two dimensions: logical coherence and medical accuracy, retaining only those that exceed the threshold in both dimensions. The filtered samples are then subjected to additional human evaluation to further ensure the quality of the data.

Figure 3 presents the data statistics of the constructed PG-Bench dataset with TBM, which consists of five subsets and

[3]https://www.youlai.cn

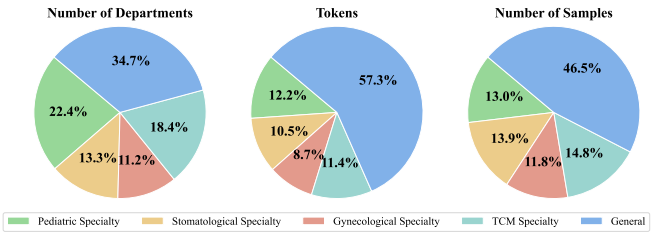

Fig. 3. Statistics of the proposed PG-Bench dataset.

contains a total of 19,814 instances across 98 departments. We split the data into training and test sets with a ratio of 7:3. Additionally, to evaluate the alignment between PG-Bench dataset and real-world clinical scenarios, we conduct a rigorous human evaluation, as detailed in Section VII-A.

C. Evaluation Metric

As defined in Section III-A, CCPG is a multi-label classification task. Therefore, we use micro-level accuracy, precision, recall, and F1 scores as automatic metrics to evaluate the performance of various methods on PG-Bench.

IV. Knowledge-Evolvable Assistant (KEA) Framework

Although LLMs have strong general capabilities, their performance is often limited on specific tasks, especially in the healthcare domain [1]. To improve performance on medical tasks, recent studies have attempted to fine-tune LLMs to acquire healthcare knowledge [43]. Nevertheless, such approaches lack the ability to support knowledge evolution.

To address these challenges and improve performance on PG-Bench, we propose a novel Knowledge-Evolvable Assistant (KEA) framework. Figure 2 illustrates the overall architecture of the KEA framework. The core module of KEA is an evolving knowledge pool (EKP), which consists of an experience bank (EB), a reflection bank (RB), and a medical external knowledge base (MEKB). Using EKP, we leverage retrieval-augmented generation to enable self-evolution of CCPG knowledge.

Specifically, given a medical dialogue, the goal of the retrieval process is to query the most relevant information from EKP. Subsequently, the medical dialogue is combined with the retrieved information to form a prompt, which is then fed into the LLMs to predict the department. For the training set, in order to continuously update the Chinese medical guidance knowledge within EKP, the predicted department generated by the LLMs is subject to label verification. Correct responses are stored in the experience bank, while incorrect responses are sent back to the LLMs for reflection generation. Afterwards, the LLMs use the medical dialogue, retrieved information, and reflections as input to predict the department again. If the prediction is correct, both the correct answer and the reflection are stored in the reflection bank. Next, we provide a detailed description of EKP and the self-evolution mechanism.

A. Evolving Knowledge Pool

The evolving knowledge pool (EKP) is designed to store external medical knowledge and dynamically collect the LLMs' accumulated experience and error-driven reflections. As shown in the lower right corner of Figure 2, EKP consists of three components: the medical external knowledge base (MEKB), the experience bank (EB), and the reflection bank (RB).

Medical External Knowledge Base: To enhance the capabilities of LLMs in domain-specific tasks without updating parameters, we design a medical external knowledge base (MEKB) to statically store knowledge related to CCPG. We integrate a wide range of medical knowledge from various sources, including online medical corpora (e.g., EBM-NLP [44] and LCMDC [45]), medical datasets (e.g., MIMIC-III [46]), academic literature (e.g., PubMed), encyclopedias (e.g., Wikipedia), clinical records and textbooks, into MEKB for subsequent retrieval. To ensure data quality and inspired by HuatuoGPT-II [47], we apply a multi-stage preprocessing pipeline including (1) domain-specific extraction using medical lexicons from THUOCL and UMLS SPECIALIST Lexicon, (2) text segmentation and normalization, (3) filtering of low-quality or advertising content using a trained quality classifier, and (4) de-duplication based on sentence embeddings and dense retrieval similarity. The resulting knowledge base provides a large-scale, clean, and semantically diverse medical foundation for subsequent retrieval-augmented reasoning.

Experience Bank: To enable LLMs to mimic human doctors in clinical practice by leveraging previous similar cases to assist in handling new ones, we design an experience bank (EB) to store past successful CCPG cases. The experience bank is structured in the form of question-answer pairs, where the question details the medical dialogues and instructions, and the answer contains the validated predicted departments.

Reflection Bank: To enable LLMs to learn from failures, we design a reflection bank (RB) to store corrected erroneous cases after reflection, along with the self-summarized reasons for the mistakes. Similar to the experience bank, the reflection bank is structured in the form of question-answer pairs, where the question details the medical dialogues and instructions, and the answer contains the validated predicted departments along with the self-summarized reasons for the errors.

B. Self-Evolution via Retrieval-Augmented Generation

The self-evolution process is accompanied by the continuous updating of medical knowledge within EKP, which is enabled by retrieval-augmented generation. As shown on the right side of Figure 2, the dynamic updating of EKP is based on the evolution of the experience bank and reflection bank. Each time the LLMs predict departments, correct answers are stored in the experience bank, while incorrect answers undergo reflection. Once the model predicts the correct departments again, both the answer and the reflection are stored in the reflection bank. The overall process consists of two parts: knowledge retrieval and dynamic evolution.

Knowledge Retrieval: Given a medical conversation D, we integrate it into the prompt instructions of the medical guidance task to form the prompt P. To effectively retrieve the external knowledge most relevant to P from the evolving knowledge pool, we construct a retriever that encodes P into a dense vector representation.

TABLE I
ZERO-SHOT PERFORMANCE COMPARISON OF LLMS ACROSS DIFFERENT SPECIALTY DOMAINS. THE "ALL" COLUMN REPRESENTS THE UNION OF THE FIRST FIVE CATEGORIES. THE BEST-PERFORMING METRICS ARE HIGHLIGHTED IN **BOLD AND UNDERLINED**. MODELS MARKED WITH * ARE COMMERCIAL APIS.

Model	Pediatric		Stomatological		Gynecological		TCM		General		All	
	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)
Mistral-7B-Instruct-v0.3 [48]	0.29	0.87	1.47	4.39	1.76	2.80	0.75	4.70	2.15	1.94	0.98	2.03
Baichuan2-7B-Chat [49]	1.01	16.53	2.93	24.12	2.24	11.53	2.00	13.40	2.36	15.65	1.99	15.92
Llama-3.1-8B-Instruct [3]	1.29	35.43	2.93	38.49	5.29	44.89	5.74	26.13	3.01	43.23	3.56	40.08
Yi-1.5-9B-Chat [50]	2.16	30.85	4.27	35.88	4.01	47.52	8.85	28.13	9.96	44.77	6.98	40.48
GLM4-9B [51]	11.65	38.48	9.87	37.81	20.51	50.00	**11.72**	33.18	21.71	48.82	16.79	44.62
GLM4-9B-chat [51]	2.01	39.71	3.20	39.43	4.97	47.26	8.35	30.37	10.69	47.96	7.47	43.63
Gemma2-9B-instruct [52]	8.78	41.18	12.93	39.96	20.03	52.01	7.11	27.05	17.76	48.26	14.88	44.31
Qwen2-7B [53]	11.08	13.61	16.27	25.61	15.54	27.17	7.48	24.20	23.46	35.99	16.99	29.13
Qwen2-7B-instruct [53]	11.37	33.63	11.07	36.17	17.63	39.66	7.61	25.19	16.46	47.17	13.52	40.68
Qwen2.5-7B-instruct [42]	10.65	42.64	11.60	38.96	16.99	51.14	10.22	32.00	15.37	50.64	13.58	45.96
Qwen2.5-72B-Instruct* [42]	12.52	46.35	14.80	40.29	17.31	51.06	9.85	32.14	18.09	50.63	15.21	46.64
DeepSeek-R1-Distill-Qwen-32B* [5]	15.68	45.35	12.00	36.79	21.63	53.25	10.60	32.53	17.52	47.67	16.36	44.91
BioMistral-7B [54]	1.45	14.23	0.93	8.67	1.76	27.43	0.62	9.27	2.85	9.29	1.48	10.51
Taiyi-7B [13]	13.53	15.32	3.47	4.31	0.00	5.61	3.62	4.15	3.78	6.34	4.47	6.56
Llama3-OpenBioLLM-8B[4]	2.30	25.61	4.53	33.32	4.49	42.44	5.36	20.86	4.72	26.82	4.71	28.72
Llama-3.1-8B-UltraMedical [55]	0.00	11.63	0.80	25.41	0.80	9.95	4.01	16.41	1.18	10.68	1.43	13.19
Lingdan-13B [16]	15.82	16.20	9.47	24.16	6.57	27.64	5.99	17.28	13.99	17.67	10.45	21.05
Sunsimiao-7B[5]	7.91	35.97	9.73	34.91	21.31	35.71	6.73	23.82	11.71	40.31	10.52	36.75
HuatuoGPT-o1-7B [10]	9.35	41.28	11.33	31.49	12.98	41.29	9.98	31.65	15.73	37.60	13.00	37.47
WiNGPT2-Gemma-2-9B-Chat[6]	23.60	33.30	24.00	38.09	30.13	49.61	5.74	26.42	23.33	41.12	21.95	39.92
BianCang-7B [12]	**33.81**	37.83	23.33	32.96	**32.37**	40.95	10.22	30.28	**27.36**	44.57	**25.02**	40.05
BianCang-7B-instruct [12]	32.37	35.98	**29.47**	36.80	24.20	33.65	7.98	27.41	25.93	41.06	23.75	37.34
DeepSeek-V3* [56]	7.91	45.60	13.07	40.46	18.59	53.32	10.10	32.90	18.54	**51.55**	15.45	**47.54**
GPT-3.5-turbo* [2]	15.11	38.39	16.13	38.94	23.50	48.71	11.51	32.63	15.08	41.42	14.87	41.08
GPT-4o* [2]	23.45	**46.83**	21.87	**40.50**	28.76	**55.57**	11.64	**34.10**	24.16	49.42	22.06	46.58

Specifically, we adopt BGE [57], a dual-encoder model with strong performance across domains as the retriever. The passage encoder is used to index the text in the evolving knowledge pool to support retrieval. At runtime, the query encoder maps the prompt P into an embedding, and similarity scores are computed to retrieve the $Top\text{-}k$ entries from the medical external knowledge base, experience bank, and reflection bank, thereby augmenting the CCPG knowledge.

Notably, the retrieved results from the three knowledge banks are hierarchically fused: experiential knowledge offers reference examples; reflective knowledge provides error corrections and reasoning strategies; and external medical knowledge supplements authoritative medical information. The fused content is subsequently structured and incorporated into the prompt, yielding the augmented prompt P_{aug}. The specific structure and format of P_{aug} are illustrated in Figure 4.

Dynamic Evolution: We input the augmented prompt P_{aug} into the LLMs to generate the response r. To acquire LLMs' experience accumulation and error-driven reflections and dynamically store them in the evolving knowledge pool, we validate each response r_i generated from the training set, handling correct and incorrect responses, respectively. (1) Correct Response: For the i-th case in the training set, if the response r_i generated by the LLMs matches the target department label y_i, we combine the prompt P_i with r_i into a structured question-answer pair (P_i, r_i) and store it in the experience bank. (2) Incorrect Response: For the i-th case in the training set, if the response r_i generated by the LLMs does not match the target department label y_i, we use P_i, r_i, and y_i as input to guide the LLMs in reflecting and generating an error summary s_i. Then, we feed P_i and s_i back into the LLMs to obtain a revised response r_i'. Next, if r_i' matches y_i, we combine the prompt P_i with r_i, r_i', and s_i into a structured question-answer pair $(P_i, r_i \oplus r_i' \oplus s_i)$, where $\oplus$ denotes concatenation.

Prompt Template for KEA

#System Prompt:
You are an intelligent medical assistant. Please identify the appropriate department for patient registration based on the provided doctor-patient dialogue and the given list of candidate departments.
#Output Requirements:
1. Output must only contain items from the candidate department list, no additional explanations.
2. The answer may contain one or more departments, separated by "|".
#Task:
Based on the doctor-patient dialogue information, recommend appropriate registration departments from the given candidate list.
#Candidate Department List: {department_list}
#Doctor-Patient Dialogue: {dialogue}
#Retrieved Contextual Information:
1. Retrieved from experience bank: {context1}
2. Retrieved from reflection bank: {context2}
3. Retrieved from medical external knowledge base: {context3}
#Output Format:
Recommended Departments: [Department_1] | [Department_2]...
#Reference: {reference}

Fig. 4. Prompt template used for inference in the KEA framework.

C. Prompt Templates

Figure 4 illustrates the prompt templates employed for inference within the KEA framework. These templates guide

the LLMs in processing medical dialogues and generating appropriate responses, ensuring structured interaction with the evolving knowledge pool. Additional prompt templates can be found in our GitHub repository[7].

V. Experiments

A. Baselines

To comprehensively evaluate the performance of various LLMs on the CCPG benchmark, we selected widely used and state-of-the-art (SOTA) LLMs from both general and biomedical domains for experimentation.

B. Experimental Details

We divide the experiments into two groups: investigating the performance of LLMs on PG-Bench and exploring the effectiveness of the KEA framework, respectively.

For LLMs, all local models (non-API) are deployed using the Swift framework [58], with the temperature to 0, max_tokens to 1024, timeout to 30 seconds, and max_retries to 3. For in-context examples (few-shot), we use the BGE retriever [57] to select the most similar data from the training set. For the KEA framework, we retrieve one piece of knowledge (k=1) from the medical external knowledge base, experience bank, and reflection bank. Additionally, we conducted experiments where we retrieved one piece of knowledge from the medical external knowledge base, and three pieces from the experience bank and reflection bank, respectively.

The experiments are conducted on a server equipped with eight NVIDIA GeForce RTX 3090 GPUs.

VI. Results and Discussion

A. Performance Evaluation of Existing LLMs on PG-Bench

Table I presents the performance of 25 distinct LLMs on PG-Bench in a zero-shot setting. According to Table I, we have several observations.

First, the latest models, such as GPT-4o and BianCang-7B, exhibit significantly better performance on PG-Bench compared to older models like Mistral-7B. Specifically, BianCang-7B achieves the highest accuracy (Acc.) across multiple specialties, surpassing both general-purpose LLMs and other biomedical-domain models. Meanwhile, GPT-4o demonstrated the best F1 scores in Stomatology (40.50%), Gynecology (55.57%), and TCM (34.10%). In contrast, Mistral-7B struggles with accuracy (Acc.) across all subsets, with the best performance being only 2.15%. This potentially indicates that the latest architectures and updated data significantly impact the performance of LLMs on task-specific scenarios.

Second, general-purpose LLMs fine-tuned with dialogue data showed a decline in accuracy (Acc.). For instance, in the "All" column, the accuracy (Acc.) of GLM4-9B-chat (7.47%) is lower than that of GLM4-9B (16.79%). This can likely be attributed to alignment issues, where the fine-tuned versions introduced model safety reinforcement learning, leading to the forgetting of knowledge from other domains.

Third, the performance of biomedical-domain LLMs is not always superior to that of general-purpose LLMs. For example, in the General column, BioMistral-7B achieves an accuracy (Acc.) of 2.85%, which is significantly lower than the 21.71% accuracy (Acc.) by GLM4-9B. This may be due to the limited exposure to Chinese medical dialogue data during the training process, despite pre-training on medical data, which results in struggles with performance on PG-Bench.

Fourth, API-based large-parameter LLMs did not show a clear advantage over smaller locally deployed LLMs. For instance, in the "All" column, GPT-4 achieves an accuracy (Acc.) of 22.06%, which is not significantly better than the 25.02% accuracy (Acc.) achieved by locally deployed LLMs like BianCang-7B. This suggests that the inclusion of high-quality domain-specific data might enable smaller parameter LLMs to achieve competitive performance on the task, even surpassing API-based LLMs in some cases.

Fifth, it is worth noting that among relatively larger-parameter LLMs, Qwen2.5-72B-Instruct achieves a marginally higher F1 score in the "All" column (46.64%) compared to DeepSeek-R1-Distill-Qwen-32B (44.91%), while the latter demonstrates better accuracy in the Gynecological (21.63% vs. 17.31%) and "All" (16.36% vs. 15.21%) columns. These results suggest that architectural design and training methodologies may play a more critical role than model scale alone in the context of CCPG tasks.

Finally, when comparing different subsets, we observed that the pediatric subset exhibited lower performance metrics. For example, in the case of GLM4-9B, the accuracy (Acc.) in the Pediatric column is 11.65%, compared to 20.51% in the Gynecological column. This can likely be attributed to the fact that the pediatric subset has a larger number of target department categories (for example, as shown in Figure 3, pediatric has more department categories than gynecology), making the task more challenging.

Overall, the performance of all LLMs on the PG-Bench was unsatisfactory, struggling across the board. The best-performing LLM achieved an ACC and F1 score of only 25.02% and 47.54% in the "All" column, which highlights the wide research opportunities for the proposed Chinese conversation-based patient guidance task and underscores the room for future improvements.

B. Performance Evaluation of the Proposed KEA Framework

Tables II, III, and IV present the performance of the KEA framework applied to three LLMs: GPT-4o, DeepSeek-V3, and BianCang-7B-Instruct. We summarize our observations below.

First, our KEA framework outperforms other model configurations, as it not only leverages rich external clinical knowledge but also learns from past experiences and reflections. For instance, as shown in Table II, in the "All" column, 1-round-KEA achieves 31.09% accuracy (Acc.) on GPT-4o (vs. 22.06% for Zero-shot).

[4]https://huggingface.co/aaditya/OpenBioLLM-Llama3-8B.
[5]https://github.com/X-D-Lab/Sunsimiao.
[6]https://github.com/winninghealth/WiNGPT2.
[7]https://github.com/KJOrigin/PG-Bench.

TABLE II
PERFORMANCE COMPARISON OF VARIOUS METHODS ON **GPT-4O** ACROSS DIFFERENT SPECIALTY DOMAINS. "N-ROUND-KEA" REFERS TO KEA APPLIED WITH N ROUNDS OF INFERENCE ON THE TRAINING SET. ※ INDICATES THE EXPERIENCE BANK AND REFLECTION BANK EACH RETURN 3 DOCUMENTS, AND THE MEDICAL EXTERNAL KNOWLEDGE BASE RETURNS 1 DOCUMENT.

Model	Method	Pediatric		Stomatological		Gynecological		TCM		General		All	
		Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)
GPT-4o	Zero-shot	23.45	46.83	21.87	40.50	28.76	55.57	11.64	34.10	24.16	49.42	22.06	46.58
	Few-shot	26.33	47.77	25.20	40.41	27.07	55.40	12.52	34.58	24.12	51.40	24.07	48.30
	1-round-KEA	29.93	51.62	24.93	43.95	36.50	**60.57**	13.58	36.37	38.13	62.54	31.09	55.32
	2-round-KEA	31.80	51.88	26.00	43.73	34.41	58.53	12.55	35.34	41.18	65.39	32.67	56.46
	3-round-KEA	31.37	52.92	24.27	44.26	33.55	59.29	13.34	35.84	41.26	65.29	32.97	56.53
	3-round-KEA※	**37.84**	**54.80**	**27.87**	**44.66**	**36.99**	60.39	**14.32**	**38.20**	**46.82**	**68.38**	**37.91**	**59.49**

TABLE III
PERFORMANCE COMPARISON OF VARIOUS METHODS ON **DEEPSEEK-V3** ACROSS DIFFERENT SPECIALTY DOMAINS.

Model	Method	Pediatric		Stomatological		Gynecological		TCM		General		All	
		Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)
DeepSeek-V3	Zero-shot	7.91	45.60	13.07	40.46	18.59	53.32	10.10	32.90	18.54	51.55	15.45	47.54
	Few-shot	8.06	46.53	13.73	40.68	16.83	53.99	10.22	32.56	18.82	50.61	15.42	46.87
	1-round-KEA	13.67	50.61	15.33	42.46	20.03	56.95	11.85	36.08	36.42	64.81	25.70	55.93
	2-round-KEA	13.53	49.86	14.27	43.13	18.59	57.00	11.97	36.27	39.19	66.84	26.02	56.40
	3-round-KEA	12.81	49.98	14.80	43.64	17.63	56.26	11.60	35.97	39.35	66.90	26.24	56.99
	3-round-KEA※	**20.29**	**52.86**	**19.43**	**45.20**	**21.96**	**57.56**	**16.21**	**38.66**	**46.20**	**70.11**	**32.34**	**59.63**

TABLE IV
PERFORMANCE COMPARISON OF VARIOUS METHODS ON **BIANCANG-7B-INSTRUCT** ACROSS DIFFERENT SPECIALTY DOMAINS.

Model	Method	Pediatric		Stomatological		Gynecological		TCM		General		All	
		Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)
BianCang-7B-instruct	Zero-shot	32.37	35.98	29.47	36.80	24.20	33.65	7.98	27.41	25.93	41.06	23.75	37.34
	Few-shot	33.53	36.28	23.47	31.21	27.40	38.24	8.48	26.56	26.30	41.70	20.32	32.27
	1-round-KEA	39.98	41.39	30.40	37.10	39.26	46.94	39.15	40.15	35.73	50.06	36.62	45.60
	2-round-KEA	**40.72**	**42.81**	31.20	37.62	40.22	47.55	42.14	41.74	37.76	51.79	37.54	46.50
	3-round-KEA	40.29	42.50	**32.67**	**39.10**	40.71	47.28	**43.77**	**44.38**	38.78	52.83	**38.15**	47.29
	3-round-KEA※	37.84	40.13	29.87	35.80	**41.03**	**49.14**	38.28	39.98	**40.81**	**55.48**	37.91	**47.89**

Second, to rule out the possibility that the performance improvement brought by the KEA framework is due to providing in-context examples to the LLMs, we conducted additional experiments in a few-shot setting. The results indicate that the impact of providing in-context examples on performance is minimal. For GPT-4o, as shown in Table II, the accuracy gap between Zero-shot and Few-shot in the "All" column is only 2.01 points. In contrast, 1-round-KEA achieves 31.09% accuracy, significantly outperforming Few-shot (24.07%). These results clearly demonstrate that the KEA framework's performance gains are not merely due to in-context examples but rather its ability to leverage external knowledge and iterative learning.

Third, to investigate the impact of the number of evolution rounds on the performance of the KEA framework, we conducted experiments with 1 to 3 rounds. Taking BianCang-7B-instruct as an example, Table IV reports accuracy under the "TCM" column: 3-round-KEA achieves 43.77% accuracy on GPT-4o, compared to 39.15% and 42.14% for 1-round-KEA and 2-round-KEA, respectively. The results indicate that as the number of evolution rounds increases, the KEA framework's performance generally improves across the majority of LLMs.

Finally, to evaluate the impact of the number of external knowledge retrievals in the KEA framework on model performance, we increased the retrieval counts in both the experience bank and reflection bank from 1 to 3. Taking DeepSeek-V3 as an example, Table III shows that across all subsets, under the same 3-round KEA setting, 3-round-KEA※ consistently achieves higher performance than 3-round-KEA. These results confirm that increasing experience and reflection during inference significantly improves LLM performance.

Overall, the KEA framework effectively enhances the performance of LLMs on PG-Bench, with capabilities improving further after each additional inference round. However, the absolute performance remains constrained by the inherent challenges of CCPG, attributed to the high number of candidate departments, complex and frequent multi-department combinations, and the long-tail distribution of less common specialties, which together create classification difficulty.

VII. ADDITIONAL EXPERIMENTS

A. Human Evaluation

To validate the quality of PG-Bench dataset, we invited three clinical professionals to annotate the data. Specifically, we randomly selected 150 dialogue samples and paired them with 150 real-world clinical dialogues from the same department labels. Each annotator rated the dialogues on six dimensions: *Fluency*, *Consistency*, *Integrity*, *Interaction Strategy*, *Safety*,

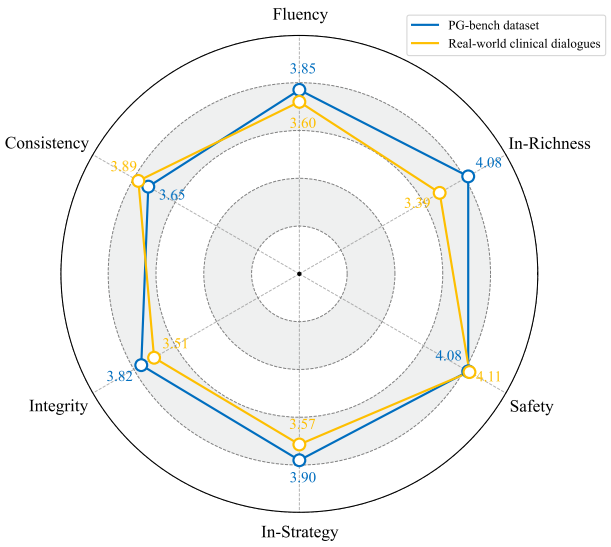

Fig. 5. Comparison of average evaluation scores on six dimensions for PG-Bench dataset versus real-world clinical dialogues. In-Strategy: *Interaction Strategy*. In-Richness: *Information Richness*.

and *Information Richness*, using a scale from 1 to 5. We calculated a Fleiss' κ score of 0.68, indicating a good level of agreement among the annotators. The annotation results are shown in Figure 5, where it can be observed that the PG-Bench dataset performs comparably to real-world clinical dialogues across all dimensions.

B. Evaluation of Fine-Tuning

To assess the effectiveness of LoRA [59] fine-tuning on the CCPG task, we performed LoRA fine-tuning of the BianCang-7B-instruct using the PG-Bench dataset and compared its performance with the base model under zero-shot and few-shot settings. As shown in Figure 6, fine-tuning significantly improved classification accuracy and F1-score on the General and "All" datasets. For example, in the General dataset, accuracy increased from 25.93% and 26.30% for zero-shot and few-shot base models respectively to 36.75% and 33.21% after fine-tuning. Correspondingly, F1-score rose from 41.06% and 41.70% to 53.55% and 49.39%. These results demonstrate that LoRA fine-tuning effectively enhances generalization across datasets and overall predictive performance.

Building upon the fine-tuned model, the KEA framework was further applied to exploit iterative knowledge-enhanced reasoning. Starting from the fine-tuned zero-shot baseline, multiple rounds of KEA consistently improved accuracy and F1-score. For instance, in the "All" dataset, F1-score increased from 53.57% after one round of KEA to 55.88% after three rounds. The General dataset also showed notable gains, with accuracy increasing from 47.56% to 49.43% and F1-score from 62.13% to 64.34%. These findings confirm that the KEA framework further enhances the fine-tuned model by effectively leveraging external knowledge.

C. Evaluation of Distribution Shift and Retrieval Robustness

To investigate the impact of data distribution differences within the Experience Bank and Reflection Bank components

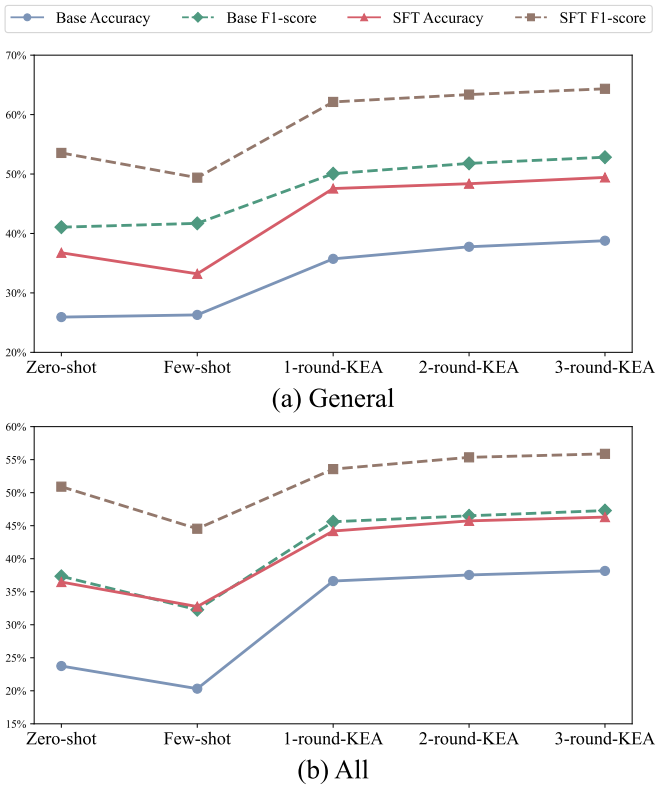

Fig. 6. Performance comparison of various methods using the Base LLM (**BianCang-7B-Instruct**) and the SFT LLM (obtained via **LoRA SFT** on PG-Bench dataset) across different specialty domains.

of the KEA framework on model performance, we conducted experiments using DeepSeek-V3, with results presented in Table V. We first analyzed the PG-Bench dataset distribution, finding that the top five departments, namely Internal Medicine, Pediatrics, Surgery, General Practice, and Rehabilitation Medicine, account for approximately 48% of the data. This pronounced skew indicates a dominance of high-frequency departments alongside many low-frequency ones. Such imbalance raises concerns regarding model performance across departments with varying representation. To quantify this effect, we constructed the Experience Bank and Reflection Bank exclusively from high-frequency department data and evaluated the model on samples from both high- and low-frequency departments. Details of the selected dataset are provided in Table VI. Results indicate that performance on low-frequency departments was not necessarily inferior; notably, accuracy on low-frequency samples exceeded that on high-frequency samples by 3.50%, with recall and F1 scores also remaining largely comparable. This suggests that the KEA framework enables the model to generalize beyond dominant distributions, mitigating risks associated with data imbalance.

In addition to data distribution, the quality and relevance of retrieved knowledge within the KEA framework's Experience and Reflection Banks play a critical role in enhancing model predictions. We further investigated the retrieval outcomes in low-frequency department scenarios. Interestingly, even when the knowledge retrieved corresponded to a different department label than the ground truth, the model frequently generated the correct prediction. This was in contrast to the vanilla baseline model, which often failed under similar conditions. This phenomenon aligns with prior research [60] showing that

indirectly related but contextually relevant documents, when properly integrated, can improve retrieval-augmented model accuracy. Although the retrieved knowledge may not perfectly match the target label, its semantic relevance appears sufficient to guide the model towards correct inferences. This robustness in knowledge retrieval and utilization underlines the strength of the KEA framework in leveraging noisy or partially relevant information to boost overall task performance.

D. Impact Analysis of Model Scales

To investigate zero-shot performance across different parameter scales on CCPG tasks, we evaluated four recently released Qwen3 models—Qwen3-4B, Qwen3-8B, Qwen3-32B, and Qwen3-235B-A22B [4]—on the PG-Bench dataset. All models were assessed under a consistent inference setting with the official *no_think* mode enabled. We report Accuracy, Precision, Recall, and F1 scores, summarized in Table VII.

The results indicate that the Qwen3-32B model achieves the best overall performance across most medical domains, attaining the highest F1 scores and demonstrating improved generalization with increased model scale up to this point. In contrast, the largest Qwen3-235B-A22B model does not consistently outperform the Qwen3-32B model, suggesting diminishing returns or potential optimization issues at extreme scale in zero-shot settings, warranting further investigation. The smaller models (Qwen3-4B and Qwen3-8B) demonstrate significantly inferior performance in comparison.

E. Ablation Study

To investigate the specific contributions of KEA's components, we performed ablation experiments on the Pediatrics subset using GPT-4o, systematically evaluating the individual and combined effects of the Experience Bank, the Reflection Bank, and the Medical External Knowledge Base.

Specifically, we systematically removed each module and their combinations to quantify their individual and joint impact on model performance. In the experiments, the Experience Bank and Reflection Bank each return 3 documents, Medical External Knowledge Base returns 1 document.

The experimental results, summarized in table VIII, clearly show that the Experience Bank is the most critical module, with its removal causing a sharp drop in accuracy (from 37.84% to 22.19%) and F1-score (from 54.80% to 49.96%), demonstrating the importance of case-based analogical reasoning in CCPG tasks. In terms of performance contribution, the Experience Bank provides the most significant performance boost, while the Reflection Bank mainly improves recall by encouraging self-correction on initially incorrect predictions, while the Medical External Knowledge Base offers modest but consistent support by enhancing factual reliability.

Moreover, configurations combining two components generally perform better than single modules alone, and the full system integrating all three achieves the best overall performance. These findings confirm that KEA's modular design yields complementary benefits and substantively contributes to performance gains.

F. KEA Performance in More Rounds

To assess the long-term stability of the knowledge evolution mechanism, we conducted extended experiments on the Pediatric subset using DeepSeek-V3, evaluating the performance from rounds 4 to 7. In the experiments, the Experience Bank, Reflection Bank and Medical External Knowledge Base each return 1 document.

As shown in Table IX, the results demonstrate that KEA maintains stable and reliable performance in multi-turn self-evolution. From the fourth round onward, F1 scores consistently exceed 50%, with the highest score of 51.01% observed in round 5. Following this peak, the model's performance begins to converge, with F1 scores fluctuating between 50.58% and 50.62% across rounds 6 and 7. Accuracy also reaches its maximum of 14.89% in round 5. These results suggest that after the initial performance gains in the early rounds, KEA enters a stable phase, maintaining reliable reasoning capability and demonstrating long-term effectiveness and stability.

Given the static nature of the training data, performance gains tend to converge with increased evolution rounds; however, in real-world clinical environments with continual data influx, KEA is expected to yield further improvements.

G. Error Analysis

To gain deeper insights into the models' failure modes, we conducted an error analysis on 200 misclassified cases from the PG-Bench test set.

Most Common Failures: Approximately 72% of the errors fall into the following two categories: (1) Multi-department predictions. In interdisciplinary cases, the model often predicts some, but not all, correct departments, indicating a limited understanding of cross-departmental dependencies. (2) Long-tail misclassification. Low-frequency departments are frequently misclassified as more common, symptomatically related departments due to data imbalance.

Most Dangerous Failures: Approximately 1% of the sampled cases involve misclassifications that could lead to inappropriate or delayed care, such as assigning *Chest Pain* to the *Surgery Department* rather than *Internal Medicine Department*, or classifying *Pediatric Convulsions* under *Child Growth and Development Specialty* instead of *Pediatric Neurology Specialty*. Although rare, such errors could have implications for patient safety.

In addition, we conducted a qualitative analysis of errors addressed by KEA and those that persisted. The self-evolution mechanism effectively corrects errors caused by isolated knowledge gaps, such as misattributing a distinct symptom to a superficially similar department. For instance, a case of *toothache and sensitivity to cold* was initially misclassified as *Prosthodontics Specialty*, but was correctly guided to *Endodontics Specialty* after KEA retrieved a reflection that clarified their distinct specializations. Despite KEA's success in correcting isolated knowledge gaps, certain errors persist across multiple evolution rounds, consistent with the common failure modes noted earlier: (1) misclassification of long-tail departments, (2) cases with ambiguous, cross-department symptoms, and (3) errors in complex multi-turn dialogues.

TABLE V
PERFORMANCE COMPARISON OF **DEEPSEEK-V3** ON HIGH-FREQUENCY AND LOW-FREQUENCY TEST SETS.

Model	Method	High-frequency Test Set				Low-frequency Test Set			
		Acc.(%)	Pre.(%)	Rec.(%)	F1.(%)	Acc.(%)	Pre.(%)	Rec.(%)	F1.(%)
DeepSeek-V3	Zero-shot	12.00	51.11	50.36	50.73	17.00	45.45	53.25	49.05
	Few-shot	11.50	49.62	47.94	48.64	19.50	47.11	52.66	50.50
	1-round-KEA	19.00	55.05	52.78	53.89	21.00	47.53	54.14	50.62
	2-round-KEA	**21.50**	55.74	56.42	56.08	21.50	46.97	57.40	51.66
	3-round-KEA	20.00	55.71	**56.66**	**56.18**	20.00	47.13	58.23	52.12
	3-round-KEA※	20.50	**56.17**	55.93	56.17	**24.00**	**47.58**	**58.28**	**52.39**

TABLE VI
DATASET STATISTICS USED FOR THE DISTRIBUTION SHIFT EXPERIMENTS. THE THREE DATASETS CONTAIN MUTUALLY EXCLUSIVE SAMPLES.

Dataset	# of Samples	Description
Training Set	600	70% from high-frequency departments
High-frequency Test Set	200	70% from high-frequency departments
Low-frequency Test Set	200	∼5% from high-frequency departments
Total	1,000	–

TABLE VII
PERFORMANCE COMPARISON OF **QWEN3** LLMS ACROSS DIFFERENT SPECIALTY DOMAINS.

Model	Pediatric		Stomatological		Gynecological		TCM		General		All	
	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)
Qwen3-4B	3.31	38.77	8.00	38.08	9.29	48.52	9.48	30.56	11.34	47.97	9.00	43.55
Qwen3-8B	2.30	36.10	3.47	37.92	6.09	46.89	9.10	31.33	4.97	40.18	5.30	39.32
Qwen3-32B	**16.98**	**46.08**	**17.07**	**41.33**	19.23	52.27	10.47	33.65	**17.56**	**51.51**	**16.36**	**47.44**
Qwen3-235B-A22B*	13.24	45.13	13.47	40.74	**19.71**	**54.27**	**10.72**	**33.98**	16.42	50.14	15.10	46.98

TABLE VIII
ABLATION STUDY OF **KEA** ON THE **PEDIATRICS** SUBSET USING **GPT-4O**. EB: *Experience Bank*. RB: *Reflection Bank*. MEK BASE: *Medical External Knowledge Base*.

Method	Acc.(%)	Pre.(%)	Rec.(%)	F1.(%)
KEA (ours)	**37.84**	**53.17**	**56.55**	**54.80**
w/o MEK Base	37.02	53.06	55.69	54.50
w/o RB	36.98	52.83	54.13	53.47
w/o EB	22.19	44.59	56.21	49.96
w/o RB + MEK Base	36.71	52.79	53.80	53.29
w/o EB + MEK Base	24.89	46.54	55.43	51.40
w/o EB + RB	30.12	44.48	50.14	47.14

TABLE IX
MORE ROUNDS PERFORMANCE OF **KEA** ON THE **PEDIATRIC** SUBSET USING **DEEPSEEK-V3**.

Method	Acc.(%)	Pre.(%)	Rec.(%)	F1.(%)
Zero-shot	7.91	38.56	55.79	45.60
Few-shot	8.06	39.34	56.93	46.53
1-round KEA	13.67	44.36	58.92	50.61
2-round KEA	13.53	43.79	57.87	49.86
3-round KEA	12.81	43.71	58.35	49.98
4-round KEA	14.41	44.78	58.25	50.64
5-round KEA	**14.89**	**45.08**	**59.13**	**51.01**
6-round KEA	14.56	44.74	58.16	50.58
7-round KEA	14.53	44.69	58.35	50.62

In summary, while the KEA framework alleviates part of these issues through contextual retrieval and reflective reasoning, challenges remain in modeling fine-grained departmental relationships, extended clinical reasoning, and ensuring safe, reliable recommendations.

VIII. CONCLUSION

In this paper, we introduce the Chinese conversation-based Patient Guidance (CCPG) task, which aims to navigate patients to the appropriate departments based on conversations in Chinese between patients and hospital staff. To benchmark the performance of various methods on CCPG, we present PG-Bench, the first evaluation benchmark designed for the CCPG task, which includes five subsets with a total of 19,814 dialogues across 98 departments. To better address the challenges of CCPG, we further propose a Knowledge-Evolvable Assistant (KEA) framework, which elaborates a self-evolution mechanism to dynamically update knowledge and enhance the guidance ability of LLMs. Experimental results show that KEA achieves state-of-the-art performance on PG-Bench, but there is still significant room for improvement, highlighting the vast potential for future research in this challenging task.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No.62376130), Program of New Twenty Policies for Universities of Jinan (No.202333008), Shandong Talent Introduction Program, and Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2025ZDZX01).

CLINICAL RELEVANCE AND IMPACT

The clinical relevance and impact of our work can be viewed from several perspectives. First, by introducing PG-Bench,

the first comprehensive benchmark for Chinese conversation-based patient guidance (CCPG), we provide a standardized and rigorous evaluation suite that reveals the limitations of current large language models (LLMs).

Second, the proposed Knowledge-Evolvable Assistant (KEA) framework demonstrates a practical pathway toward improving real-world CCPG systems. By integrating validated experiences, error-driven reflections, and external medical knowledge, KEA substantially enhances LLM performance without costly fine-tuning. This iterative knowledge evolution mirrors the learning process of human clinicians and holds potential to reduce misclassification risks in patient guidance.

Moreover, the potential clinical benefits are significant. Automated CCPG systems could alleviate the burden on front-line hospital staff, streamline patient navigation, and shorten waiting times, ultimately improving access to timely and appropriate care. In particular, the lightweight and scalable design of KEA enables its widespread deployment across diverse hospital environments.

Finally, the open-source release of PG-Bench and KEA ensures accessibility, reproducibility, community-driven validation, and continuous improvement. This democratization of patient guidance research fosters collaboration across academia, industry, and healthcare institutions, accelerating the translation of AI-driven solutions into clinical practice.

Limitations

While our KEA framework demonstrates promising performance, it is inherently built upon LLMs, which are known to exhibit hallucination—i.e., generating factually incorrect or misleading information. This limitation poses potential risks in high-stakes domains such as healthcare, where factual accuracy is critical. Mitigating hallucinations remains a challenging open problem and will be a focus of our future work.

Our KEA framework lies in its reliance on the retrieval mechanism to dynamically augment the knowledge pool. Although the retriever is intended to identify relevant external knowledge, it may inadvertently introduce erroneous or irrelevant information into the evolving knowledge base.

While the template-based method enables scalable and privacy-safe dialogue generation, it may oversimplify real patient–clinician interactions. The reliance on predefined templates and LLM outputs can introduce linguistic bias and reduce the authenticity and diversity of conversational behaviors.

During dynamic evolution, cases that remain incorrect after reflection are discarded to maintain the overall quality of stored experiences. While this strategy effectively ensures the reliability of the reflection bank, it may also slightly bias the retained data toward more typical or easier cases. In future work, we plan to log such persistently challenging cases for targeted analysis and incremental learning.

Ethical Considerations

Although our work focuses on improving patient guidance with large language models, the deployment of these models in high-stakes domains such as healthcare may introduce risks, including inherent biases and limitations in clinical reasoning.

Given that patient dialogues often involve sensitive or personally identifiable information, we implemented strict measures to ensure that the datasets we constructed and utilized were strictly confined to the scope of the research and contained no sensitive personal data.

References

[1] X. Liu, H. Liu, G. Yang, Z. Jiang, S. Cui, Z. Zhang, H. Wang, L. Tao, *et al.*, "A generalist medical language model for disease diagnosis assistance," *Nature Medicine*, vol. 31, no. 1, pp. 932–942, 2025.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, pp. 1–100, 2023.

[3] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, pp. 1–92, 2024.

[4] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, pp. 1–35, 2025.

[5] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, *et al.*, "Deepseek-R1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, pp. 1–22, 2025.

[6] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao, *et al.*, "Large AI models in health informatics: Applications, challenges, and the future," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 12, pp. 6074–6087, 2023.

[7] Y. Wu, K. Mao, Y. Zhang, and J. Chen, "CALLM: Enhancing clinical interview analysis through data augmentation with large language models," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 12, pp. 7531–7542, 2024.

[8] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[9] Y. Tian, R. Gan, Y. Song, J. Zhang, and Y. Zhang, "ChiMed-GPT: A chinese medical large language model with full training regime and better alignment to human preferences," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 7156–7173, 2024.

[10] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, "HuatuoGPT-o1, towards medical complex reasoning with LLMs," *arXiv preprint arXiv:2412.18925*, pp. 1–23, 2024.

[11] L. Sun, D. Liu, M. Wang, Y. Han, Y. Zhang, B. Zhou, Y. Ren, *et al.*, "Taming unleashed large language models with blockchain for massive personalized reliable healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 6, pp. 4498–4511, 2025.

[12] S. Wei, X. Peng, Y. Wang, T. Shen, J. Si, W. Zhang, F. Zhu, A. V. Vasilakos, W. Lu, X. Wu, *et al.*, "Biancang: A traditional Chinese medicine large language model," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2025.

[13] L. Luo, J. Ning, Y. Zhao, Z. Wang, Z. Ding, P. Chen, W. Fu, Q. Han, G. Xu, Y. Qiu, *et al.*, "Taiyi: A bilingual fine-tuned large language model for diverse biomedical tasks," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1865–1874, 2024.

[14] Z. Zhao, W. Lu, X. Peng, L. Xing, W. Zhang, and C. Zheng, "Automated ICD coding via contrastive learning with back-reference and synonym knowledge for smart self-diagnosis applications," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 6042–6053, 2024.

[15] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, and H. Zan, "Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue," in *Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence*, pp. 19368–19376, 2024.

[16] R. Hua, X. Dong, Y. Wei, Z. Shu, P. Yang, Y. Hu, S. Zhou, H. Sun, K. Yan, X. Yan, *et al.*, "Lingdan: Enhancing encoding of traditional Chinese medicine knowledge for clinical reasoning tasks with large language models," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 2019–2029, 2024.

[17] Y. Luo, J. Zhang, S. Fan, K. Yang, M. Hong, Y. Wu, M. Qiao, and Z. Nie, "BioMedGPT: An open multimodal large language model for biomedicine," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2024.

[18] W. Lu, S. Wei, X. Peng, Y.-F. Wang, U. Naseem, and S. Wang, "Medical question summarization with entity-driven contrastive learning," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 4, pp. 1–19, 2024.

[19] B. Cao, S. Huang, and W. Tang, "AI triage or manual triage? Exploring medical staffs' preference for AI triage in China," *Patient Education and Counseling*, vol. 119, no. 1, pp. 1–7, 2024.

[20] S. Paslı, A. S. Şahin, M. F. Beşer, H. Topçuoğlu, M. Yadigaroğlu, and M. İmamoğlu, "Assessing the precision of artificial intelligence in emergency department triage decisions: Insights from a study with ChatGPT," *The American Journal of Emergency Medicine*, vol. 78, no. 1, pp. 170–175, 2024.

[21] M. Lu, B. Ho, D. Ren, and X. Wang, "TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage," in *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing*, pp. 5747–5764, 2024.

[22] J. Li, Y. Lai, W. Li, J. Ren, M. Zhang, X. Kang, S. Wang, P. Li, Y.-Q. Zhang, W. Ma, *et al.*, "Agent hospital: A simulacrum of hospital with evolvable medical agents," *arXiv preprint arXiv:2405.02957*, pp. 1–29, 2024.

[23] D. Peta, A. Day, W. S. Lugari, V. Gorman, V. M. T. Pajo, *et al.*, "Triage: A global perspective," *Journal of Emergency Nursing*, vol. 49, no. 6, pp. 814–825, 2023.

[24] E. Dippenaar, "Triage systems around the world: A historical evolution," *International Paramedic Practice*, vol. 9, no. 3, pp. 61–66, 2019.

[25] J.-P. Gaudilliere, A. McDowell, C. Lang, and C. Beaudevin, "Triage beyond the clinic," *Global Health for All. Knowledge, Politics, and Practices*, pp. 78–101, 2022.

[26] Priyanka and D. Kumar, "Decision tree classifier: A detailed survey," *International Journal of Information and Decision Sciences*, vol. 12, no. 3, pp. 246–269, 2020.

[27] L. Jiang, D. Wang, Z. Cai, and X. Yan, "Survey of improving naive bayes for classification," in *Proceedings of the 3rd Advanced Data Mining and Applications*, pp. 134–145, 2007.

[28] L. Jiang, Z. Cai, D. Wang, and S. Jiang, "Survey of improving k-nearest-neighbor for classification," in *Proceedings of the 4th international conference on fuzzy systems and knowledge discovery*, pp. 679–683, 2007.

[29] J. Cervantes, F. Garcia-Lamont, L. Rodriguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, no. 1, pp. 189–215, 2020.

[30] A. B. Shaik and S. Srinivasan, "A brief survey on random forest ensembles in classification model," in *Proceedings of the 2nd International Conference on Innovative Computing and Communications*, pp. 253–260, 2019.

[31] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 1937–1967, 2021.

[32] B. Lindemann, T. Muller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia Cirp*, vol. 99, no. 1, pp. 650–655, 2021.

[33] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, 2021.

[34] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, no. 1, pp. 111–132, 2022.

[35] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: A review of BERT-based approaches," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.

[36] S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan, S. Ding, J. Wang, K. Xu, Y. Fang, *et al.*, "Large language models for disease diagnosis: A scoping review," *arXiv preprint arXiv:2409.00097*, pp. 1–69, 2024.

[37] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," *arXiv preprint arXiv:2401.06775*, pp. 1–22, 2023.

[38] P. C. Sukhwal, V. Rajan, and A. Kankanhalli, "A joint LLM-KG system for disease Q&A," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 3, pp. 2257–2270, 2025.

[39] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, *et al.*, "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, no. 1, pp. 943–950, 2025.

[40] X. Wang, G. Chen, S. Dingjie, Z. Zhiyi, Z. Chen, Q. Xiao, J. Chen, F. Jiang, J. Li, X. Wan, *et al.*, "CMB: A comprehensive medical benchmark in Chinese," in *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6184–6205, 2024.

[41] S. Macherla, M. Luo, M. Parmar, and C. Baral, "Mddial: A multi-turn differential diagnosis dialogue dataset with reliability evaluation," *arXiv preprint arXiv:2308.08147*, pp. 1–10, 2023.

[42] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, pp. 1–26, 2024.

[43] D. Yuan, E. Rastogi, G. Naik, S. P. Rajagopal, S. Goyal, F. Zhao, B. Chintagunta, and J. Ward, "A continued pretrained LLM approach for automatic medical note generation," in *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1–7, 2024.

[44] B. Nye, J. J. Li, R. Patel, Y. Yang, I. J. Marshall, A. Nenkova, and B. C. Wallace, "A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 197–207, 2018.

[45] X. Wang, H. Li, D. Zheng, and Q. Peng, "Building a Chinese medical dialogue system: Integrating large-scale corpora and novel models," *arXiv preprint arXiv:2410.03521*, pp. 1–9, 2024.

[46] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, *et al.*, "MIMIC-III, A freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[47] J. Chen, X. Wang, K. Ji, A. Gao, F. Jiang, S. Chen, H. Zhang, S. Dingjie, W. Xie, C. Kong, *et al.*, "HuatuoGPT-II, One-stage training for medical adaption of LLMs," in *Proceedings of the First Conference on Language Modeling*, pp. 1–31, 2024.

[48] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, pp. 1–9, 2023.

[49] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan, *et al.*, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, pp. 1–28, 2023.

[50] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, G. Wang, H. Li, J. Zhu, J. Chen, *et al.*, "Yi: Open foundation models by 01.AI," *arXiv preprint arXiv:2403.04652*, pp. 1–26, 2024.

[51] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao, *et al.*, "ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools," *arXiv preprint arXiv:2406.12793*, pp. 1–19, 2024.

[52] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, pp. 1–21, 2024.

[53] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, pp. 1–26, 2024.

[54] Y. Labrak, A. Bazoge, E. Morin, P.-a. Gourraud, M. Rouvier, and R. Dufour, "Biomistral: A collection of open-source pretrained large language models for medical domains," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 5848–5864, 2024.

[55] K. Zhang, S. Zeng, E. Hua, N. Ding, Z.-R. Chen, Z. Ma, H. Li, G. Cui, B. Qi, X. Zhu, *et al.*, "Ultramedical: Building specialized generalists in biomedicine," *Advances in Neural Information Processing Systems*, vol. 37, no. 1, pp. 26045–26081, 2024.

[56] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, *et al.*, "DeepSeek-V3 technical report," *arXiv preprint arXiv:2412.19437*, pp. 1–53, 2024.

[57] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, "C-Pack: Packed resources for general Chinese embeddings," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024.

[58] Y. Zhao, J. Huang, J. Hu, X. Wang, Y. Mao, D. Zhang, Z. Jiang, Z. Wu, B. Ai, A. Wang, *et al.*, "Swift: A scalable lightweight infrastructure for fine-tuning," in *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pp. 29733–29735, 2025.

[59] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proceedings of The 18th International Conference on Learning Representations*, pp. 1–26, 2022.

[60] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, and F. Silvestri, "The power of noise: Redefining retrieval for rag systems," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729, 2024.