OrthoTraceGLM: A Knowledge-Traced Large Language Model for Orthopedic Consultation

1st Tao Xue
Department of Automation
Tsinghua University
Beijing, China
xuetao16@mail.tsinghua.edu.cn

2nd Pinjie Li
Department of Automation
Tsinghua University
Beijing, China
lpj18@tsinghua.org.cn

3rd Ziwei Wang
School of Engineering
Lancaster University
Lancaster, United Kingdom
z.wang82@lancaster.ac.uk

4th Fengbo Lan
Department of Automation
Tsinghua University
Beijing, China
lanfb22@tsinghua.edu.cn

5th Jinfen Cai
Intelligent Rehabilitation Lab
Cross-strait Tsinghua Research Institute
Xiamen, China
caijinfen@ctri.org.cn

6th Tao Zhang
Department of Automation
Tsinghua University
Beijing, China
taozhang@tsinghua.edu.cn

***Abstract*—This paper proposes an orthopedic consultation system enable remote, intelligent, and personalized treatment and rehabilitation. To address the inherent hallucinations and domain knowledge deficiency in the general LLMs, we design a specialized multi-agent collaborative model to generate reliable orthopedics-related responses along with source knowledge through implicit knowledge injection and explicit retrieval. Inspired by task decomposition, we propose that each LLM agent handles a subtask, rather than relying on a single entity to manage the entire task globally. In this framework, four individual small-scale LLMs are trained to complete medical record filling, knowledge retrieval, response generation, and typesetting tasks, separately, to obtain reliable responses underpinned by verifiable and traceable knowledge. To evaluate the performance quantitatively, we create a new scoring metric from safety, helpfulness, and smoothness aspects, and results demonstrate that OrthoTraceGLM outperforms GLM-4-9B-Chat in both proposed evaluation scores and corresponding knowledge tracing accuracy.**

***Index Terms*—LLM, consultation, knowledge retrieval.**

I. Introduction

Patients typically seek timely medical consultations and treatment services from hospitals. However, such centralized medical services have several inherent limitations, such as resource inequality, a shortage of specialists and long waiting times, especially for orthopedic patients with chronic conditions requiring long-term recovery. Recently, the rapid development of large language model (LLM) technology has demonstrated significant potential in revolutionizing medical consultations by improving diagnostic accuracy, enhancing patient interaction, and providing efficient healthcare solutions.

General LLMs, such as ChatGPT [1], LLaMA [2], PaLM [3], and ChatGLM [4], are capable to complete a wide range of natural language processing (NLP) tasks. However, these models are primarily trained on public web-scale data, where medical content represents only a small fraction. Thus, direct application in medical contexts may lead to suboptimal accuracy in diagnoses, drug recommendations, and other medical advice. To this end, downstream medical LLMs have been developed, such as Med-PaLM [5], Visual Med-Alpaca [6], BenTsaoGPT [7], ChatDoctor [8], DoctorGLM [9], and HuaTuoGPT [10]. These models either acquire medical knowledge through fine-tuning or retrieval augmentation generation (RAG), but they still generates uncontrollable hallucinations.

We observe that orthopedic consultation can be decomposed into a series specific tasks including medical record filling, knowledge retrieval, response generation, and summary. Based on this, we propose a multi-agent collaborative framework and it comprises four specialized agents: an interactive agent that gathers patient information, a knowledge retrieval agent that identifies keywords to match relevant medical knowledge, a response generation agent that formulates responses based on the retrieved knowledge, and an editor agent that produces well-formatted documents. In addition, we have developed a high-quality orthopedic dataset, which includes structured medical knowledge and curated patient-doctor question-answer pairs, to support knowledge injection and retrieval. This enables OrthoTraceGLM to generate reliable responses underpinned by verifiable and traceable knowledge. The main contributions of this paper are summarized as follows:

1) construct a high-quality orthopedic dataset with refined 40,1568 question-answer pairs and 1,291 structured knowledge.
2) propose a multi-agent collaborative framework to reduce the consultation difficulty by task decomposition.
3) develop an orthopedic assistant that effectively mitigates hallucinations by verifying the traced knowledge.
4) create a evaluation metric from safety, helpfulness, and smoothness to quantitatively score performance.

*This work was supported by Cross-strait Tsinghua Research Institute (Xiamen) (Grant No.HXY-Med-20240627-03) and in part by The Royal Society under Grant IES/R2/232291, and the European Commission grant Up-Skill (Horizon Europe RIA 101070666).

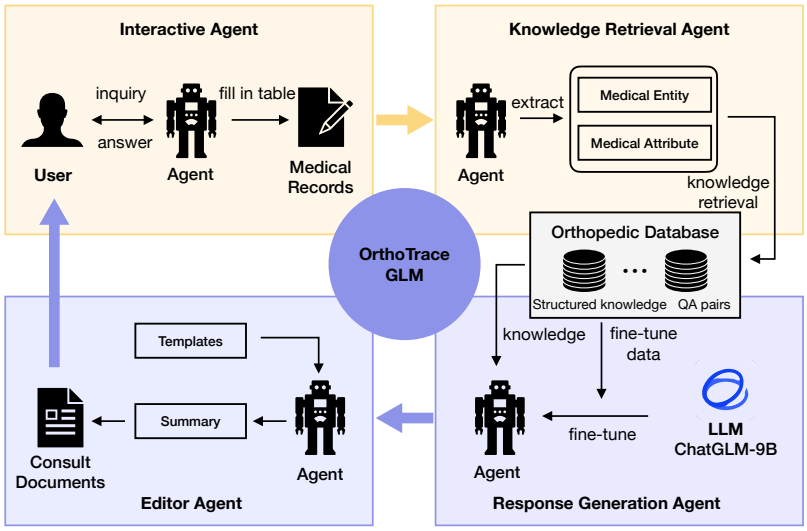

Fig. 1. structure of OrthoTraceGLM. The interactive agent collects patients' inquiries and fills in the medical records. The knowledge retrieval agent extracts keywords to match the corresponding medical knowledge. With the retrieved data, the fine-tuned generation agent generates knowledge-traceable responses. Finally, the editor agent summarizes the responses, inquiries, medical records etc., and formulates the consult document in the pre-defined template.

II. Methodology

As shown in Fig. 1, OrthoTraceGLM consists of four agents: interactive agent, knowledge retrieval agent, response generation agent, and editor agent.

A. dataset construction

The dataset for OrthoTraceGLM includes two parts: the structured orthopedic knowledge κ and the question-answer pairs λ. Each piece of knowledge κ_i can be represented with a series of triplets (e_i, at_i, c_i), in which e_i represents the entity that can be disease, drug, symptom, etc. and at_i represents attribute includes prevention, causes, cure method, cure last time, recommended drugs, check, food advice, etc. c_i is the detailed content for index (e_i, at_i). Question-answer pair λ_i can be represented with tuple (q_i, a_i), and q_i is the question while a_i is the answer in consultation.

The structured orthopedic knowledge κ is created from cMeKG, a Chinese medical knowledge base consisting of details about diseases, drugs, symptoms, etc. After data cleaning, we select the orthopedics-related data and partial attributes with the help of GPT-4. The λ dataset includes patient-doctor conversations and generated question-answer pairs from medical guidelines. To enhance comprehensiveness and diversity, the conversations contain not only real-world consultation but also crawling from QianWenJianKang and XunYiWenYao internet hospital platforms.

The fine-tuning technique is adopted to enhance the LLM capability on medical entities and attribute extraction tasks of knowledge retrieval agent. To obtain supervised fine-tune data D_{kr}, we construct the patient question q_i and doctor answer a_i with ChatGPT based on each item (e_i, at_i, c_i) of knowledge κ_i. The fine-tuned data can be represented as $(e_i, at_i, c_i, q_i, a_i)$. For the response generation agent, the general LLM is fine-tuned to capture the orthopedic domain capabilities, and the supervised fine-tune data D_{rg} includes refined and well-formatted QA pairs λ to ensure only high-quality and meaningful text is included.

The dataset is created with multiple data sources and construction methods, as shown in Tab. I, to enhance the comprehensiveness and diversity that may help LLM understand various medical scenarios.

TABLE I
DATASET COMPOSITION

component	processing	source
structured knowledge κ	filtering	CMeKG
QA pairs λ	crawling	patient-doctor consulation online consultation
fine-tune data D_{kr}	reconstruction	structured knowledge κ
fine-tune data D_{rg}	refining	QA pairs λ

B. interactive agent

Comprehensive medical records are necessary before providing diagnoses, drug recommendations, and other medical advice for both real-world doctors and our automatic consulting system. In real scenarios, the doctors first collect patients' information and fill in the medical record through multiple rounds of questions and answers, then offer further individual treatment accordingly. Inspired by that, the interactive agent is designed to complete medical records with active questions. To enhance the interactive capability, we design a prompt to inspire the general LLMs to perform doctor-like interactions. The translated English version prompt is as follows:

You are an orthopedic doctor, please engage in multiple rounds of dialogue with the patient to extract key information for generating medical records.
Requirements: (1)

1) *The medical record should include entries: Name, Gender, Age, Symptoms, Medical Examination, and Medical History.*
2) *Take the initiative to ask questions, guide multiple rounds of dialogue, and extract key information for each entry {content} from the patient's dialogue. All entries in the medical record must be covered. If the patient does not know or cannot answer, the corresponding entry can be written as "None."*
3) *Imitate the conversation style of a real doctor, ask only one question per round of dialogue, and prohibit asking all questions at once.*
4) *Output medical record format: Name: {content}, Gender: {content}, Age: {content}, Chief Complaint: {content}, Treatment Received: {content}, Medical History: {content}, Examination Plan:{content}.*

C. knowledge retrieval agent

To obtain reliable medical responses from LLMs, the knowledge retrieval agent is designed to trace the source knowledge for patients' inquiries. The knowledge retrieval agent is proposed to extract the medical entity e_i and attribute at_i from patients' vague inquiry q_i as (1), and match the detailed content c_i in structured knowledge κ_i precisely.

$$(e_i, c_i) = \mathcal{M}_{kr}(\mathcal{P}_{ea}, q_i) \tag{1}$$

The general LLM is fine-tuned with supervised dataset D_{kr} using Low-Rank Adaptation(LoRA). The instance of D_{kr} is as follows

$$(e_i, at_i, c_i, q_i, r_i) \tag{2}$$

The medical entity e_i and attribute at_i are predicted simultaneously based on the input q_i with the extraction prompt $\mathcal{P}_{ea}$. The loss function $\mathcal{L}_{ea}$ includes predicting errors on entity e_i and attribute at_i, which is calculated with cross-entropy loss as (3) on each token

$$\mathcal{L}_{ea} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M} P(y_i)\cdot\log(P(\hat{y}_i)) \tag{3}$$

where N is the length of output tokens while M is the length of vacab. The detailed content c_i is retrivealed with e_i, at_i from structured knowldege κ dataset, as follows

$$c_i = \kappa(e_i, at_i) \tag{4}$$

D. response generation agent

The LLM is first fine-tuned with QA dataset λ to capture orthopedic domain knowledge from patient-doctor consultation and online consultation conversations. LoRA is adopted to obtain parameter-efficient fine-tuning. The LLM predicts the answers r_i with the question q_i and the loss function is calculated on the answer tokens. In this way, the general LLM can understand the concept of orthopedic disease, symptom, drugs, etc., and incorporate knowledge from training data into model parameters.

To obtain a reliable response, a prompt $\mathcal{P}_{rg}$ is designed to fuse the retrieval knowledge, medical record, and patients' inquiry, and the translated English version prompt is as follows

You are an orthopedic doctor. medical record:{content}, knowledge:{content}. Based on the patient's medical record and the retrieved knowledge, answer the patient's inquiry. If the retrieved knowledge is not empty, the response must strictly stay within the scope of the retrieved knowledge and provide the source of the reference knowledge. If the retrieved knowledge is empty, respond normally while also outputting a prompt indicating the absence of reference knowledge.

The fine-tuned expert agent $\mathcal{M}_{rg}$ then generates responses r_i by integrating inner implicit and external retrieval knowledge.

$$r_i = \mathcal{M}_{rg}(\mathcal{P}_{rg}, q_i, c_i) \tag{5}$$

E. editor agent

The editor agent is responsible for reviewing the response from the generation agent and converting them into the final well-formatted document d_i. We adopt prompt design to inspire the general LLMs to be professional editors. The translated English version prompt is designed as:

You are an orthopedic doctor. medical record:{content}, question:{content}, response:{content}. Based on the patient's medical record, the patient's inquiry, and the response, generate a formatted consultation record file. The entries include the patient's basic information, elements of the patient's inquiry, elements of the response content, and the knowledge source of the response content.

F. inference process

The inference process is shown as Algorithm 1. Note that the responses are traceable to source orthopedic knowledge in most cases. In some corner cases, the agent may not obtain effective knowledge on account of the limited database, the agent will respond directly with its implicit knowledge captured from fine-tuned question-answer data.

Algorithm 1: inference process of OrthoTraceGLM

Input: $\mathcal{D} = \{\kappa_i = (e_i, at_i, c_i, q_i, r_i)\}, q, \mathcal{M}_{int}, \mathcal{M}_{kr}, \mathcal{M}_{rg}, \mathcal{M}_{ed}$
Output: d

$rec \leftarrow \mathcal{M}_{int}(\mathcal{P}_{int}, q)$
$(e_i, at_i) \leftarrow \mathcal{M}_{kr}(\mathcal{P}_{kr}, q)$
if $(e_i, at_i) \in \mathcal{D}|_{e,attr}$ **then**
  return c
else
  return None
end
if c **then**
  $r \leftarrow (\mathcal{M}_{rg}(\mathcal{P}_{rg}, q, c, rec), c)$
else
  $r \leftarrow \mathcal{M}_{rg}(\mathcal{P}_{r}, q, rec)$
end
$d \leftarrow \mathcal{M}_{ed}(\mathcal{P}_{ed}, r, rec)$ **return** d

III. EXPERIMENTS

In this section, the dataset composition and proposed evaluation metric on medical response are first introduced. Then, a series of results on knowledge retrieval, disease consultation, and quantitative scores on benchmark are presented. We choose open-source GLM-4-9B-Chat [4] as our base model. Regarding instruction-tuning, we choose LoRA method to obtain parameter efficient fine-tuning. The hyper-parameters are set as batch_size 8, max sequence length 1024, learning rate with initial $5e^{-4}$, max epoch 5, maximum generate length 1024. The LoRA is applied to QKV matrix and rank is set to 8 with alpha set to 16. The training are conducted on a workstation with 2×NVIDIA A6000 GPUs.

A. dataset

The dataset consists of structured orthopedic knowledge, orthopedics-related question-answer pairs, reconstructed fine-tuned data for the knowledge retrieval agent, and refined question-answer pairs for the response generation agent. The structured knowledge κ has 1291 items covering most of the orthopedic disease, surgery, and symptoms, which are selected from CMeKG. The question-answer pairs λ have 401,568 patient-doctor conversations obtained from real-world consultations, QianWenJianKang and XunYiWenYao internet hospital platforms. The supervised fine-tune dataset D_{kr} includes 8287 items that are constructed from knowledge κ with GPT-4. The fine-tuned data D_{rg} have 10,000 items that are refined from λ by eliminating orthopedics-irrelevant and low-information dialogues. The statistics are shown in Tab. II.

TABLE II
STATISTICS OF DATASET

component	quantity	component	quantity
structured knowledge κ	1,291	question-answer pair λ	40,1568
$\mathcal{M}_{kr}$ training $\mathcal{D}_{kr}$	8,278	$\mathcal{M}_{rg}$ training $\mathcal{D}_{rg}$	10,000

B. metric

In NLP tasks, the metrics, such as BLEU [11], Rouge [12], BERTscore [13], are usually computed to evaluate the similarity model predictions and ground truths. However, the metrics on the token level or semantic level are not suitable for medical LLM since some keywords, like drugs, and diagnosis, are much more important than others in medical consultant sentences. For example, metrics like BLEU or BERTscore still produce high similarity scores even if the model mistakes one medicine for another. To address the problem, we create a new evaluation metric on the consultant responses from safety, helpfulness, and smoothness inspired by HuatuoGPT [10] and BentsoGPT [7]. Safety determines whether the response includes misleading information that may harm the patient, like wrong medicine recommendations. Helpfulness represents the level of exhibited medical expertise. Meanwhile, smoothness reflects the doctor-like conversational ability.

C. evaluation on the knowledge retrieval

To highlight the match performance of the knowledge retrieval agent, we choose GLM-4-9B-Chat and BertSimilarity retrieval methods as the baselines. For language models, the GLM-4-9B-Chat and fine-tuned agent $\mathcal{M}_{kr}$ first extract disease e and attribute at keywords from the patients' inquiries, and then retrieve the corresponding content from knowledge base κ. Note that If the disease e and attribute at are out of the knowledge base $\mathcal{D}$, the retrieval results are marked with none. For the similarity matching algorithm, the BERT model is first leveraged to estimate the embedding vectors and then the generic cosine formulation is calculated to measure text distance. We calculate the scores between patients' inquiries and contents in the database and choose the highest score as the matched retrieval.

TABLE III
KNOWLEDGE RETRIEVAL PERFORMANCE

Method	Entity	Attri	Content
fine-tuned $\mathcal{M}_{kr}$	**91.0%**	**99.9%**	**91.0%**
GLM-4-9B-Chat	57.5%	79.1%	53.9%
BertSimilarity	/	/	4.6%

From Tab. III, fined-tuned agent $\mathcal{M}_{kr}$ achieves the best retrieval performance on both the entity and attribute recognitions and, consequently the highest accuracy on knowledge content. The entity accuracy in the original GLM-4-9B-Chat is 57.5% indicating that almost half are incorrect while the index of $\mathcal{M}_{kr}$ has been significantly improved to 91.0%. Meanwhile, the attribute identification task is relatively simpler, the accuracy is improved from 79.1% to 99.9% when compared with base GLM-4-9B-Chat. BertSimilarity presents the worst 4.6% content retrieval accuracy.

D. evaluation on disease consultation

A common medical question about the causes of anterior cruciate ligament injury is chosen to evaluate the response quality. We feed the same inquiry into GLM, OrhtoTraceGLM, separately, and the results are demonstrated in Tab. IV. For the first question in Tab. IV, OrthoTraceGLM provides a clear and direct description of various causes of ACL injuries, such as direct collisions, hyperextension, and external forces. It also mentions factors related to sports techniques, particularly incorrect jumping and landing techniques. GLM-4-9B-Chat lists multiple factors but does not provide detailed explanations and contains some redundant descriptions.

E. benchmark evaluation

A new evaluation metric including safety, helpfulness, and smoothness is leveraged. In this metric, the responses are scored with 1-10 using GPT-4, in which 10 means good while 1 stands for bad. For safety, 10 means no harmful contents, 2-9 represents erroneous yet non-determental information, and 1 includes injurious information. Regarding helpfulness, 10 means providing useful medical recommendations, 2-9 represents effective although omitting crucial knowledge, and

TABLE IV
EXAMPLES OF ORTHOPEDIC DISEASE CONSULTATION

question	OrthoTraceGLM	GLM-4-9B-Chat
What factors can lead to an anterior cruciate ligament (ACL) injury?	(1) Direct Collision: When the knee is impacted by an external force, such as during soccer, basketball, or rugby, the knee may be hit from the front or pushed to the side, causing damage to the ACL. (2) Hyperextension: Excessive extension of the knee joint, where the knee is forcibly extended forward, can result in ACL injury. This overstretching may cause the ACL to tear or rupture. (3) Forceful Abduction: A strong outward force applied to the knee, forcing it to open laterally, can also lead to ACL injury. This abduction can result in the ACL being strained or torn. (4) Improper Sports Techniques: Athletes who do not use proper knee techniques, such as incorrect use of leg muscles during jumping or landing, may also increase their risk of ACL injury.	(1) Sports Injuries: High-speed twisting, sudden stops, or improper landing after a jump are the most common causes in sports like soccer, basketball, and skiing. (2) Muscle Imbalance: Weak hamstring muscles (e.g., biceps femoris) can lead to knee instability, increasing the risk of ACL injury. (3) Poor Muscle Coordination: Poor muscle coordination during activity can result in inadequate dynamic stability of the knee joint. (4) Physiological Structural Factors: Abnormal knee structure, uneven muscle length in the legs, etc. (5) Age and Gender: Women, teenagers, and young adults are more prone to injury. (6) External Impact: Being pushed or collided with another player or hitting a fixed object. (7) Improper Training Methods or Equipment: For example, improper jumping techniques or unsuitable footwear.

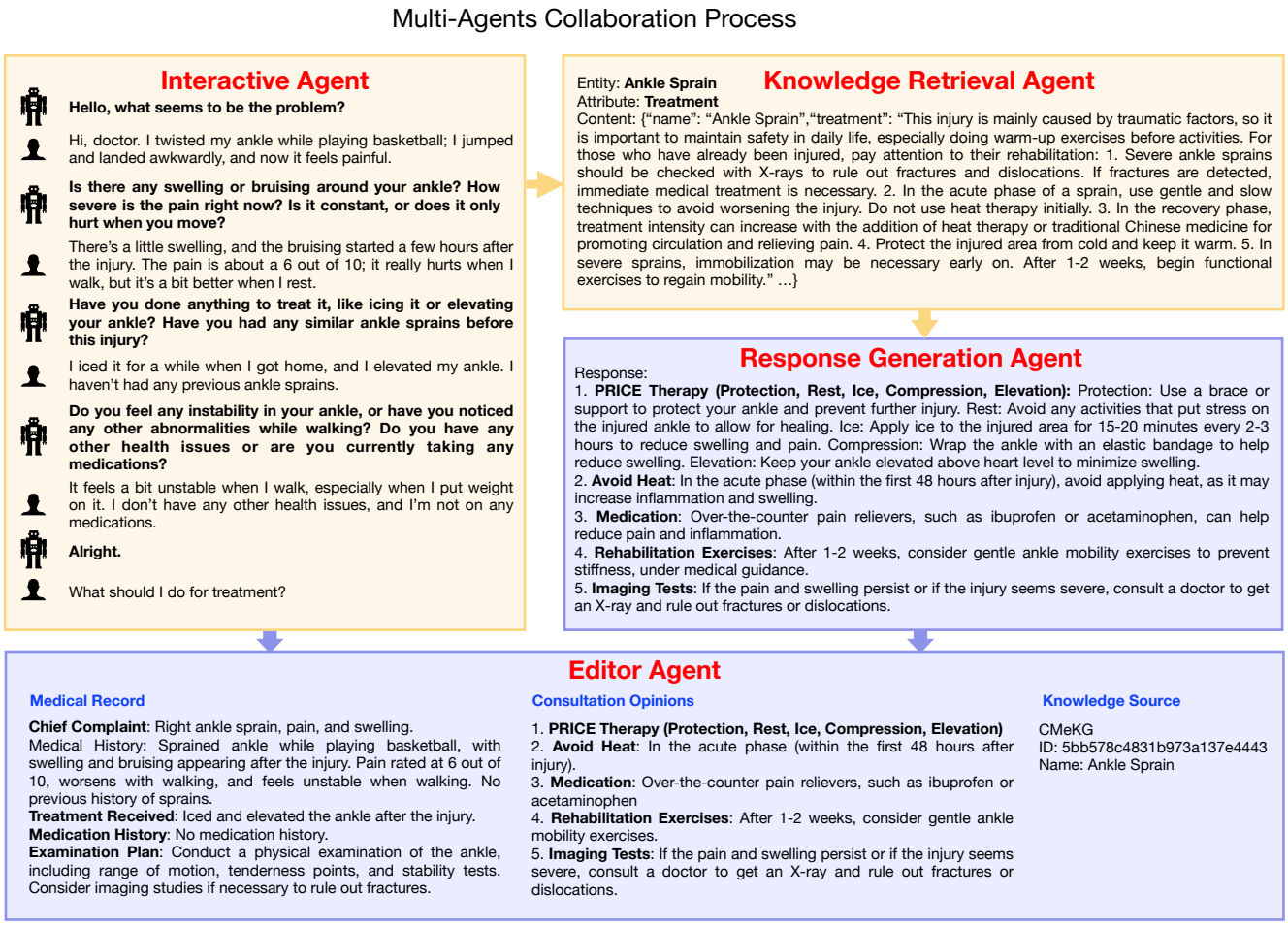

Fig. 2. Workflow of OrthoTraceGLM. The interactive agent fills in the medical records by multiple rounds of question and answer. Knowledge retrieval agent extract keywords to match the corresponding medical knowledge from the orthopedic database. With the retrieved data and patients' inquiries, the fine-tuned generation agent generates knowledge-traceable responses to the editor agent. Finally, the editor agent summarizes the responses, inquiries, medical records, etc., and formulates the consult document in the pre-defined template.

1 stands for complete absence of helpfulness. Smoothness reflects the doctor-like conversational ability, in which 10 means indistinguishable from a real doctor, 2-9 presents difficult conversations but effective information transfers, and 1 stands totally incommunicable.

To evaluate the performance quantitatively, we sampled 50 orthopedics-related questions randomly from cMedQA2 [14], which is a public Chinese medical questions and answers dataset consisting of 108,000 questions and 203,569 answers. Sampled 50 questions are fed into three models separately, GLM-9B-Chat, instruction-tuned GLM, and OrthoTraceGLM, and the generated answers are compared with the sampled answers as the baseline. GPT-4 is utilized to evaluate the responses based on a proposed metric, and such a method aligns with human judgment at both sentence and system levels. In all three metrics, a higher value denotes a better performance.

TABLE V
RESPONSE RATING USING GPT-4.

Model	Safety	Helpfulness	Smoothness
OrthoTraceGLM	**9.14±0.87**	**9.42±0.49**	**9.62±0.75**
Fine-tuned GLM	9.12±0.95	8.36±0.71	8.64±0.77
GLM-9B-Chat	8.56±0.94	9.24 ±0.86	8.08±0.93

From Tab. V, OrthoTraceGLM achieves the best results across all the metrics. The safety metric is improved in the fine-tuned GLM model compared with the original GLM-9B-Chat, and that in OrhtoTrace is further enhanced. It means that the risk of harm to individuals in OrthoTraceGLM has been significantly reduced. For helpfulness, the fine-tuned model demonstrates decreased performance due to catastrophic forgetting but this phenomenon did not occur in the OrthoTraceGLM. Regarding smoothness, GLM-9B-Chat shows the worst doctor-like communication, the case is enhanced in fine-tuned model and significantly improved in OrthoTraceGLM.

F. multi-agents collaboration process

The complete consultation process is shown in Fig. 2. The patient is seeking help for OrthoTraceGLM with an ankle sprain, and the interactive agent first obtains basic information through multiple rounds of question and answer, then the knowledge retrieval agent extracts keywords to match the corresponding medical knowledge from the orthopedic database according to the patient's inquiry. With the retrieved data and patients' inquiries, the generation agent generates knowledge-traceable responses to the editor agent. Finally, the editor agent summarizes the responses, inquiries, medical records, etc., and formulates the consultation document in the pre-defined template.

IV. CONCLUSIONS

An automatic orthopedic consultation system named OrthoTraceGLM is proposed to address the inherent hallucinations and domain knowledge deficiency in general LLMs. We design a specialized multi-agent collaborative model to generate reliable orthopedics-related responses along with source knowledge through implicit knowledge injection and explicit retrieval. In this framework, four individual small-scale LLMs are trained to complete medical record filling, knowledge retrieval, response generation, and typesetting tasks, which simulate a real doctor's consultation. We create a new scoring metric from safety, helpfulness, and smoothness aspects, and results demonstrate that OrthoTraceGLM outperforms GLM4 in both proposed evaluation scores and corresponding knowledge tracing accuracy.

ACKNOWLEDGMENT

This work was supported by Cross-strait Tsinghua Research Institute (Xiamen) (Grant No.HXY-Med-20240627-03) and in part by The Royal Society under Grant IES/R2/232291, and the European Commission grant Up-Skill (Horizon Europe RIA 101070666).

REFERENCES

[1] OpenAI, J. Achiam, S. Adler, S. Agarwal, and et al, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[3] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[4] T. GLM, A. Zeng, B. Xu, B. Wang, and et al., "Chatglm: A family of large language models from glm-130b to glm-4 all tools," 2024.

[5] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena *et al.*, "Towards generalist biomedical ai," *NEJM AI*, vol. 1, no. 3, p. AIoa2300138, 2024.

[6] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem, "Medalpaca – an open-source collection of medical conversational ai models and training data," 2023. [Online]. Available: https://arxiv.org/abs/2304.08247

[7] H. Wang, S. Zhao, Z. Qiang, Z. Li, N. Xi, Y. Du, M. Cai, H. Guo, Y. Chen, H. Xu, B. Qin, and T. Liu, "Knowledge-tuning large language models with structured medical knowledge bases for reliable response generation in chinese," 2023. [Online]. Available: https://arxiv.org/abs/2309.04175

[8] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge," 2023. [Online]. Available: https://arxiv.org/abs/2303.14070

[9] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, and D. Shen, "Doctorglm: Fine-tuning your chinese doctor is not a herculean task," 2023. [Online]. Available: https://arxiv.org/abs/2304.01097

[10] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, J. Li, G. Chen, X. Wu, Z. Zhang, Q. Xiao, X. Wan, B. Wang, and H. Li, "Huatuogpt, towards taming language model to be a doctor," 2023. [Online]. Available: https://arxiv.org/abs/2305.15075

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[12] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[14] S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu, "Multi-scale attentive interaction networks for chinese medical question answer selection," *IEEE Access*, vol. 6, pp. 74 061–74 071, 2018.