Neurocomputing 652 (2025) 131116

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

AcuGPT-Agent: An LLM-powered intelligent system for acupuncture-based infertility treatment

Jing Wen[a,1], Diandong Liu[b,1], Yuxin Xie[c,*], Yi Ren[d], Jiacun Wang[e], Youbing Xia[a,*], Peng Zhu[f,*]

[a] College of Acupuncture Moxibustion and Tuina, Nanjing University of Chinese Medicine, Nanjing, 210023, PR China
[b] Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, 710021, PR China
[c] Department of Neurosurgery, Nanjing Drum Tower Hospital, Medical School of Nanjing University, Nanjing, 211166, PR China
[d] School of Software Engineering, Nanjing University, Nanjing, 210093, PR China
[e] Department of Computer Science and Software Engineering, Monmouth University, W.Long Branch, NJ 07764, USA
[f] School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210094, PR China

ARTICLE INFO

Communicated by N. Zeng

Keywords:
AcuGPT
AcuGPT-Agent
Large language model
Acupuncture
Artificial intelligence

ABSTRACT

This paper introduces AcuGPT-Agent, a novel intelligent system powered by a domain-specific large language model (LLM) designed for acupuncture-based infertility treatment. Traditional Chinese Medicine (TCM), particularly acupuncture, presents unique challenges for general-purpose LLM due to its complex theoretical framework and specialized terminology. To bridge this gap, we make three key contributions. First, we develop AcuGPT, the first LLM specifically fine-tuned for the acupuncture domain, utilizing advanced techniques such as LoRA, DoRA, and a Pretrain + Supervised Fine-Tuning (SFT) pipeline. These methods significantly improve the ability of the model to understand and generate medical content related to TCM. Second, we propose AcuGPT-Agent, a modular and extensible system that integrates AcuGPT with standard agent tools, a domain-specific acupuncture knowledge graph for infertility treatment, and a novel LLM-driven Multi-Department Knowledge Base Routing Management Mechanism (MDKBRMM). This mechanism enhances retrieval-augmented generation by effectively managing semantic complexity and routing queries across specialized knowledge bases in medical contexts. Third, we build EvalAcu, a dedicated benchmark dataset for evaluating large language models and intelligent agents within the acupuncture domain. Experimental results demonstrate that AcuGPT exhibits strong domain knowledge and reasoning capabilities, while AcuGPT-Agent further enhances performance in real-world infertility treatment tasks. All code, datasets, and resources are publicly available at: https://github.com/suda7788/AcuGPT-Agent.

1. Introduction

Traditional Chinese Medicine (TCM) acupuncture represents a unique and systematic medical discipline. Due to its remarkable therapeutic effects, minimal side effects, and wide range of applications, acupuncture has attracted significant attention from medical communities both domestically and internationally [1,2]. In Japan, approximately 14.1 % of people have received acupuncture treatment in the last 12 months, and 25.4 % have received acupuncture at some point in their lives [3]. In the United States, acupuncture is classified as part of complementary and alternative medicine and is often used alongside conventional treatments to improve therapeutic results or alleviate symptoms [4]. In China, acupuncture is widely used in clinical treatments and health care in top-tier hospitals (Class III Grade A). As a fundamental part of TCM studies, knowledge of acupuncture is a required subject. In 2023, there were approximately 270,000 undergraduate and postgraduate students studying acupuncture [5,6].

However, the learning and transmission of acupuncture face severe challenges. The acupuncture system is extensive and profound.

* Corresponding authors.
Email addresses: wwenjingmail@163.com (J. Wen), 221612075@sust.edu.cn (D. Liu), 2450466916@qq.com (Y. Xie), jwang@monmouth.edu (J. Wang), xybd1968@njucm.edu.cn (Y. Xia), pzhu@njust.edu.cn (P. Zhu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neucom.2025.131116
Received 1 June 2025; Received in revised form 10 July 2025; Accepted 23 July 2025
Available online 25 July 2025

It encompasses a large number of specialized terms, abstract theoretical expressions, and intricate knowledge of the meridians, acupoints, and clinical skills [7,8]. The high degree of specialization, semantic abstraction, and complex structure of this knowledge system raises the barrier to learning and mastery [9]. Furthermore, since acupuncture knowledge is largely based on the reference and study of the works of renowned experts, this research approach is relatively inefficient and severely limits the speed and scale of talent cultivation, resulting in a serious shortage of professional acupuncturists [10].

With the rapid development of artificial intelligence, especially Large Language Models (LLMs), new opportunities have emerged for the modernization and transmission of acupuncture [11]. Thanks to their powerful semantic understanding and knowledge reasoning capabilities, LLMs could theoretically reduce the difficulty of learning acupuncture, improve diagnostic decision-making efficiency, and thus effectively alleviate the challenges posed by long training cycles and insufficient numbers of trained professionals [12,13]. However, several major barriers block the effective application of LLMs in the acupuncture domain. First, domain-specific data collection remains a major bottleneck. Acupuncture knowledge is often dispersed in non-standardized formats, such as classical literature, expert notes, and historical texts, making it difficult to curate structured high-quality datasets suitable for the training of LLMs [14,15]. Second, although state-of-the-art LLMs have achieved impressive results in general natural language tasks, their performance in highly specialized fields such as acupuncture remains limited. On the one hand, existing models are highly based on structured high-quality annotated data for SFT, but the unstructured nature of acupuncture knowledge, characterized by semantic ambiguity and domain-specific terminology, poses significant challenges for effective training [16,17]. On the other hand, while Retrieval-Augmented Generation (RAG) methods can incorporate external knowledge to some extent, they often fall short when dealing with content that is conceptually dense, highly specialized, and requires multilayered reasoning. Issues such as fragmented searches, incoherent context, and weak knowledge integration frequently result in outputs that lack the depth and accuracy necessary for rigorous medical decision-making [18]. Third, there is a clear lack of systematic and fine-grained evaluation tools tailored to domain-specific contexts, which hinders the effective evaluation and optimization of LLMs in practical medical applications.

To address the aforementioned challenges, we gathered a broad and comprehensive set of acupuncture-related data from multiple sources. We filtered and cleaned the data to create a high-quality pre-training dataset, integrating acupuncture knowledge into the language model. Subsequently, we utilized Deepseek to process the generated data and perform SFT on the language model, resulting in the development of AcuGPT. This specialized model adheres to the medical principles of acupuncture and enables fluent dialogue. Building on this foundation, we developed an LLM-driven framework, AcuGPT-Agent, to enable the model to serve as a proficient medical professional capable of providing accurate information. To this end, we integrated external knowledge, such as infertility acupuncture treatments, incorporated agent modules, and introduced a novel LLM-driven Multi-Department Knowledge Base Routing Management Mechanism (MDKBRMM). This mechanism enhances the system's capability to handle specialized medical knowledge complexity, enabling efficient query routing across domain-specific knowledge bases and improving performance in highly specialized tasks. To evaluate the performance of AcuGPT and the AcuGPT agent in domain-specific acupuncture consultations, we developed a customized evaluation dataset, EvalAcu. This dataset enables systematic assessment of state-of-the-art LLMs in understanding and applying professional knowledge of acupuncture. The results demonstrate the significant efficacy and practical advantages of the proposed approach in enhancing the expertise in the acupuncture domain of LLMs and the capacity to apply it in the real world.

The main contributions of this work are as follows: (1) we develop AcuGPT, the first pre-trained and fine-tuned LLM specifically designed for the acupuncture domain. (2) Based on AcuGPT, we propose AcuGPT-Agent, a modular intelligent system that integrates external knowledge with an innovative retrieval-augmented mechanism to improve the accuracy of knowledge retrieval in acupuncture-related tasks. (3) We construct EvalAcu, a comprehensive and fine-grained benchmark dataset to systematically evaluate the knowledge and task performance of LLMs and intelligent agents within acupuncture-specific contexts.

2. Related work

This section reviews previous research at the intersection of LLMs and TCM, with a specific focus on their potential for development in the treatment of acupuncture and infertility.

2.1. Large language models in traditional Chinese medicine domain

The advancement of AI, particularly in Natural Language Processing (NLP), has led to the development of powerful LLMs, such as OpenAI's GPT series and models such as Qwen2 and DeepSeekR1. These models excel in language understanding and generation, but often struggle to comprehend domain-specific terminology or perform complex reasoning [19,20]. To address these limitations, techniques such as domain-specific fine-tuning, continuous pretraining, and instruction tuning are commonly used [21–23]. These approaches help tailor LLMs for specific applications, improving their performance in tasks such as clinical dialogue and analysis.

In recent years, the application of LLMs in the medical field has gained substantial momentum, extending into areas such as TCM. Models namely HuatuoGPT [24], HuatuoGPT-II [25], MedChatZH [26], TCMChat [27], and Zhongjing [28] have been developed to incorporate TCM-specific knowledge through domain-adaptive pre-training and fine-tuning. However, the application of LLMs to acupuncture, particularly in complex therapeutic processes such as acupuncture-based infertility treatment, remains underexplored. The intricate integration of diagnosis and treatment in acupuncture, based on meridian theory and syndrome differentiation, presents challenges to current LLMs, which often lack the depth and contextual sensitivity required for such detailed clinical decision-making.

2.2. Acupuncture-based infertility treatment

Infertility is defined as the failure to conceive after one year of regular unprotected sexual intercourse with a partner [29]. In 2021, the global number of women affected by infertility reached 110.1 million, with the overall incidence continuing to increase [30]. In the United States, approximately 12.7 % of women of reproductive age seek infertility treatment each year [31]. Assisted Reproductive Technologies (ART) remain the primary treatment option. However, their application is limited by factors such as multiple invasive procedures, prolonged treatment cycles, and uncertain outcomes [32]. As a vital component of TCM, acupuncture is now practiced in 196 countries and regions around the world and has been included in health insurance systems in many countries [33,34]. Extensive research has confirmed that acupuncture is a safe and low risk therapeutic approach with the potential to regulate bodily functions and support recovery [35,36]. Since 1999, when clinical trials first demonstrated the benefits of acupuncture in improving ART outcomes [37], its use in infertility treatment has expanded significantly, gaining wide recognition from both clinicians and patients. Acupuncture has shown positive effects in improving oocyte quality, promoting ovulation, enhancing endometrial receptivity, reducing the incidence of ovarian hyperstimulation syndrome during in vitro fertilization and embryo transfer, and increasing live birth rates [38,39].

Despite its promise, the clinical application of acupuncture in infertility treatment faces several challenges. These include its highly personalized nature, the significant influence of the expertise of the physician on the outcomes [40], and the lack of standardized treatment protocols [41]. In addition, patient education resources are limited and there is insufficient knowledge exchange between healthcare providers

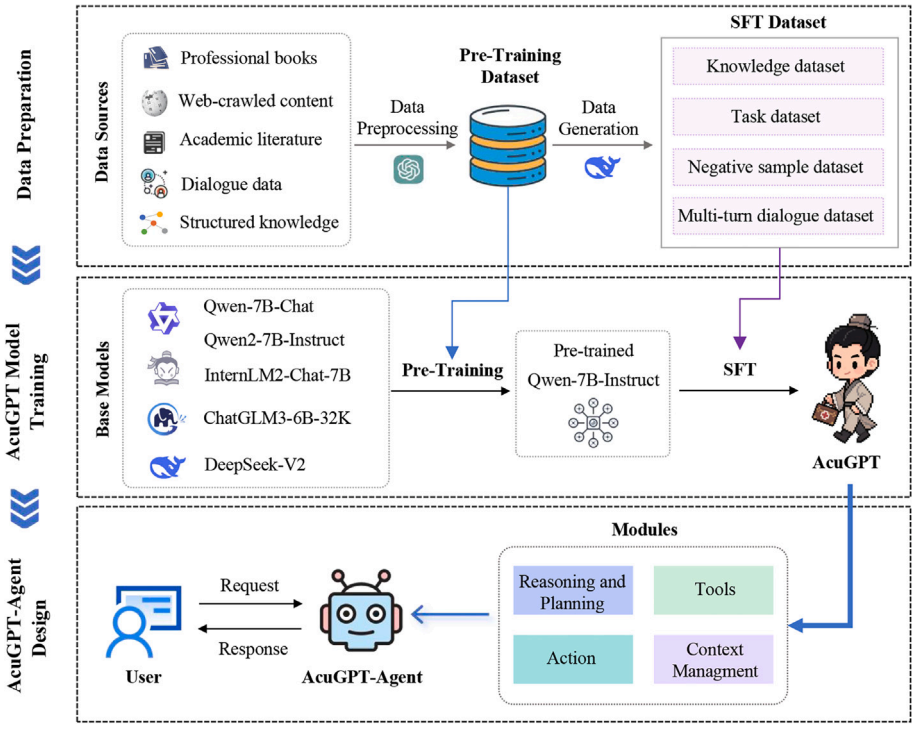

Fig. 1. The overall workflow of AcuGPT-Agent framework. Source data from books, literature, and other references is gathered and processed to create a pre-training dataset, which is used to train base models, followed by Supervised Fine-Tuning (SFT) to develop AcuGPT. AcuGPT serves as the brain of the system, interacting with users through specialized modules within the AcuGPT-Agent framework, including reasoning and planning, tools, action, and context management, ensuring dynamic responses tailored to user requirements.

[42], hindering the integration of acupuncture into the main medical systems [43]. To address these challenges, this article presents AcuGPT-Agent, the first intelligent agent system tailored for acupuncture-based infertility treatment. Built on AcuGPT, a domain-specific LLM fine-tuned with acupuncture data, it excels at understanding and generating specialized medical content. By integrating external tools, knowledge graphs, and reasoning algorithms, AcuGPT-Agent delivers robust clinical decision support and sets a new benchmark for AI-assisted acupuncture applications.

3. Methodology

To enhance the semantic understanding of LLMs in the acupuncture domain and improve response accuracy in specific clinical scenarios such as infertility treatment, we propose AcuGPT-Agent, a modular agent built on AcuGPT. The workflow of the framework is illustrated in Fig. 1.

3.1. Data sources

We collected, processed and integrated information from a wide range of high-quality data sources, covering both domain-specific and general content. The sources span book data, web-crawled content, research papers, knowledge graphs, open source datasets, and dialogue data. The content encompasses not only professional acupuncture knowledge, but also general language and instruction-following capabilities.

Book-based sources: Textual data were extracted from 10 national and industry standards, 7 widely used acupuncture textbooks adopted by higher education institutions in TCM, and 82 classical medical texts that form the theoretical and clinical foundation of acupuncture. These materials comprehensively cover acupuncture theory, techniques, diagnostics, and treatment practices. A complete list of the referenced works is provided in the appendix (see Supplementary Table S1).

Structured knowledge bases: A total of 54,593 structured records were obtained from the Acupuncture Indications Knowledge Database (ACU-IKD, http://acuai.njucm.edu.cn:8081/), which include information on acupoints, associated diseases, symptoms, and syndrome patterns. A total of 21,637 structured triples were extracted from the Ownthink open Chinese knowledge graph (http://www.ownthink.com). These triples represent essential semantic relationships that involve acupoints, meridian systems, and fundamental concepts in TCM, enriching the representation of structured knowledge.

Academic literature: Approximately 300,000 abstracts of academic articles were recovered using targeted keywords such as acupuncture, moxibustion and acupoints. Data acquisition was performed through web crawling from major academic databases, including CNKI, PubMed, and Web of Science.

Dialogue data: A collection of 2000 short-form acupuncture related questions and 1000 additional question and answer records were compiled from widely used TCM platforms and health-focused Q&A forums. These samples reflect practical clinical inquiry scenarios and serve as valuable data for training models in realistic consultation-style dialogue.

Web-crawled content: Additional domain-relevant textual material was obtained from Chinese Wikipedia and the multilingual Wikipedia corpus (retrieved on 23 January 2025 from https://dumps.wikimedia.org/), contributing to the general language modeling capability.

Open source general datasets: To enhance the generalization of the model and counteract potential degradation from fine-tuning, we incorporated 500,000 samples from publicly available open-source datasets. These datasets cover categories like logical reasoning, instruction execution, and multiturn dialogue. Detailed sources are listed in Table 1.

Table 1
General supplementary datasets.

Datasets name	Types	Description	Source
alpaca_gpt4_data	Instruction-following data	52,000 samples for enhancing instruction understanding.	https://agie.ai/datasetsdetails/alpaca-gpt4
CoT_Chinese_data	Reasoning data	Chain-of-thought data for improving logical reasoning.	https://huggingface.co/datasets/QingyiSi/Alpaca-CoT
Belle_open_source_1M	Instruction data	1 million samples for handling complex Chinese instructions.	https://huggingface.co/datasets/BelleGroup/train_1M_CN/blob/main/Belle_open_source_1M.json
multiturn_chat_0.8 M	Multi-turn dialogue data	800,000 dialogues for realistic long-context conversations.	https://huggingface.co/datasets/BelleGroup/multiturn_chat_0.8M

3.2. Data preprocessing

The large-scale acupuncture domain-specific dataset is divided into smaller sections of 1000 tokens each, with a 250-token overlap between adjacent fragments to maintain contextual continuity and prevent information loss. This can be specified in Python syntax:

$$B_i = D\,[i \times 750 \,:\, i \times 750 + 1000]\,, \tag{1}$$

where, B_i denotes the chunk whose index is i, and the 750-token stride specifies how far the window moves forward at each step, not counting the 250-token overlap. D denotes the complete token-ordered text (treated as a list or array), and the slice $i \times 750 \,:\, i \times 750 + 1000$ selects the half-open range beginning at index $i \times 750$ (inclusive) and ending just before $i \times 750 + 1000$.

For each B_i, we use LLMs (such as Deepseek-V2 and GPT-4.0) to extract knowledge, transforming complex texts into specific and concise knowledge. The knowledge extraction is represented by the following mapping function:

$$K_j = f\left(B_i\right), \tag{2}$$

where, K_j represents the j-th piece of knowledge extracted from B_i, and the f is the model extraction function.

For each K_j, the BGE-M3 model [44] is used for vectorization, mapping the knowledge to a high-dimensional vector space and deduplication of the dataset. This process can be achieved using cosine similarity:

$$\mathrm{Cos}(x, y) = \frac{x \cdot y}{\|x\|\|y\|}, \tag{3}$$

where, x and y represent the vector representations of two knowledge entries. When the similarity $\mathrm{Cos}(x, y)$ exceeds a certain threshold, they are considered duplicate data.

After removing duplicates from K_j, we use the base of the Qwen2-7 b-instruct model to regenerate concise and complete content C_k, forming a domain-specific pre-training dataset.

$$C_k = g(K_j), \tag{4}$$

where, g is the generation function of Qwen2-7 b-instruct.

3.3. Data generation

To further align AcuGPT with domain-specific tasks and dialogue scenarios in acupuncture, we adopt a dual-stage data construction strategy that processes the pre-training corpus into structured data suitable for SFT. This enables the model to effectively learn both domain-relevant knowledge and task-oriented interaction patterns.

For the knowledge-based fine-tuning dataset D_K, we use DeepSeekV2 to process the raw text T from the pre-training corpus, extracting the set of knowledge G_i. Then, based on the G_i, the corresponding question-answer pairs $K(Q_i, A_i)$ are generated. This process can be represented as:

$$D_K = \mathrm{DeepSeekV2}\left(T \rightarrow G_i \rightarrow K(Q_i, A_i)\right). \tag{5}$$

For the task-based fine-tuning dataset D_T, we first construct a diverse set of task templates $M = M_1, M_2, \ldots, M_n$ and use the DeepSeekV2 model to randomly traverse the G_i. For each traversed G_i, we randomly select a task template and strictly follow that template to generate the task-based fine-tuning dataset $T(G_i, M_j)$. This process can be represented as:

$$D_T = \sum_{i=1}^{n} T(G_i, Random(M)). \tag{6}$$

3.4. AcuGPT model training

3.4.1. Dataset

To effectively support the training of AcuGPT models, we have constructed a structured and multisource integrated training data system, which includes pre-training dataset and SFT dataset. This strategy aims to gradually equip the model with deep domain knowledge and practical task-solving capabilities. Table 2 lists the detailed information about the pre-training and SFT datasets used in AcuGPT.

For the pre-training dataset, we preprocessed acupuncture texts from books and research papers. In addition, acupuncture-specific knowledge is extracted from ACU-IKD and KnowledgeGraphData, which contain acupoints, meridians, diseases, symptoms, and syndrome patterns, etc. The final curated corpus consists of approximately 1.01 billion tokens, which form the core of the domain-specific pretraining dataset. To enhance general language understanding while preserving domain specificity, we incorporated a refined subset of Chinese Wikipedia, consisting of 122,934 high-quality articles selected from the January 23, 2025 dump. Moreover, to promote multilingual generalization and enrich semantic diversity, we integrated the Chinese and English portions of the full multilingual Wikipedia dump. All data from sources were processed using the previously described chunking, knowledge extraction, data extraction by duplicate, and deduplication. This component of the data set serves as a broad knowledge base to support the generalizability of language models beyond the acupuncture-specific domain.

For the SFT dataset, we developed a curated corpus, which comprises approximately 100,000 single-turn question-answer pairs, 3000 multi-turn dialogue samples, and 500,000 pieces of general supplementary data. This dataset was built upon the domain-specific pre-training corpus, and further enriched through collaborative annotation involving human experts and LLMs from the GPT series. All data samples were carefully reconstructed and refined to ensure factual accuracy, domain relevance, and stylistic consistency. After further processing, the final dataset was categorized into four types: knowledge dataset, task dataset, negative sample dataset, and multi-turn dialogue dataset.

The knowledge dataset was developed using a structured methodology focused on factual Q&A pairs related to the meridians, acupoints, treatment principles, and other core areas of knowledge in acupuncture. For the task dataset, each sample was organized into three parts:

Table 2
Overview of Pre-training and SFT datasets.

Data type	Content	Data structure	Size	Source
Pre-training	Professional books			Textbook, National standard
Pre-training	Triplet	Strings	1.01 B tokens	ACU-IBD, KnowledgeGraphData
Pre-training	Academic literature			CNKI, Web of Science, PubMed
Pre-training	Open source	Strings	122,934 articles	Chinese Wikipedia
Pre-training	Crawler	Strings	118,000 articles	Multilingual Wikipedia dump
SFT	Knowledge dataset	Q&A pairs	70,000 samples	Domain corpus + human-LLM collaborative annotation
SFT	Task dataset	Dictionary	30,000 samples	Human-LLM annotation with structured prompts
SFT	Negative sample dataset	Q&A pairs	>5000 samples	Knowledge dataset
SFT	Multi-turn dialogue dataset	Q&A pairs	3000 samples	TCM platforms and health-focused Q&A forums
SFT	General supplementary dataset	Dictionary	500,000 pieces	Open-source language model training dataset

"instruction" as task prompt, "input" as optional supporting information, and "output" as correct response. The instructions were generated in a variety of formats and all outputs were carefully reviewed to ensure correctness and relevance. The final samples were formatted in JSON for use in fine-tuning. The negative sample dataset was constructed by extracting questions from existing knowledge-based samples and generating responses using the model. Responses that were vague or incorrect were explicitly marked as incorrect. This method supports error-aware training and encourages the model to improve its reasoning ability. The dataset of multi-turn dialogue samples was generated by simulating expert-level discussions using the GPT-4o model. The process included identifying core themes from acupuncture knowledge and guiding the model to produce coherent, domain-relevant conversations based on those themes.

Through this process, we produced high-quality question and answer datasets that cover a wide range of practical scenarios in acupuncture, providing a robust foundation for model training.

3.4.2. Training strategy

Existing open source LLMs exhibit a limited ability to represent concepts of TCM, particularly those related to acupuncture. To enhance the adaptability of the models to acupuncture-related tasks, we initially selected five open source base models as candidates. Qwen2-7B-Chat, Qwen2-7B-Instruct, InternLM2-Chat-7B, ChatGLM3-6B-32K, and DeepSeek-V2.

In practice, to improve resource efficiency, we did not perform full domain-specific pre-training and fine-tuning on all candidate models. Instead, we first conducted a systematic evaluation of each original model's general capabilities and acupuncture domain knowledge capabilities. The goal was to identify the base model that is most sensitive to knowledge of acupuncture and demonstrates the best performance in domain tasks.

Based on the evaluation results, we selected the model that showed the strongest performance in domain-specific understanding and reasoning as the primary model for subsequent domain-specific pre-training and SFT. On this basis, we further explored different optimization strategies and fine-tuning techniques to enhance model's performance in clinical scenarios.

During the pre-training phase, we conducted incremental pre-training on the selected base model using the curated pre-training dataset, enabling the model to progressively absorb structured acupuncture knowledge, including theoretical systems, domain-specific terminology, and treatment paradigms. The training hyperparameters, such as the optimizer, learning rate, and batch size, were dynamically configured according to the following formula:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t), \tag{7}$$

where, θ_t represents the current model parameters, η is the learning rate, and $L(\theta_t)$ is the loss function. Its gradient, $\nabla_\theta L(\theta_t)$, guides the direction and magnitude of the parameter updates. This training stage optimizes the model parameters by minimizing the loss function, thereby integrating in-depth and specialized acupuncture knowledge into the base model and establishing a solid foundation for subsequent acupuncture related tasks [45].

Following the pre-training phase, we performed supervised fine-tuning using the constructed SFT dataset to further align the model with downstream clinical tasks. To explore the most effective optimization method, we investigated different training strategies, including LoRA [46], DoRA [47], and LoRA+ [48]. In addition, we evaluated two combined strategies: the Pretrain (LoRA) + SFT (LoRA) approach [49], and the Pretrain + SFT paradigm [50]. This stage enables the model to better understand and execute specific tasks within clinical scenarios, enhancing its applicability to real-world medical contexts.

3.5. AcuGPT-Agent

This section introduces AcuGPT-Agent, an intelligent and modular agent framework designed to support acupuncture based infertility treatment (Fig. 2). It integrates domain-specific acupuncture knowledge with advanced reasoning capabilities to handle both question-response and complex clinical workflows in TCM acupuncture. Its modular architecture standardizes functional components and improves tool reusability, allowing dynamic scheduling of internal tools to complete end-to-end tasks such as symptom analysis, syndrome differentiation, and acupuncture prescription generation.

The core modules of AcuGPT-Agent are the tools module and the reasoning and planning model. The tools module is a central innovation that comprises a set of customized tools tailored to various clinical scenarios in acupuncture, especially infertility treatment. These tools are supported by adaptive matching algorithms that connect patient-specific information with TCM principles, ensuring accuracy and personalized treatment. Its extensible design also allows for easy integration of new tools as the system evolves. Another key innovation is the Multi-Department Knowledge Base Routing Management Mechanism (MDKBRMM), embedded within the reasoning and planning model. This mechanism leverages LLMs to intelligently route queries across structured medical knowledge bases, facilitating efficient retrieval and cross-disciplinary reasoning.

At the core of the system is AcuGPT, which collaborates through iterative reasoning with other modules to execute complex clinical workflows. The architecture is further supported by a context management module, a foundational component responsible for maintaining task-relevant context by integrating patient history and previous interactions, thus ensuring consistency and continuity in clinical decision-making.Together, these components form a cohesive framework that unifies domain expertise, dynamic tool use, and robust reasoning.

In the following sections, we present details of the architecture and functionalities of the three core modules.

3.5.1. Tools module

To operationalize its modular design, the tools module is implemented through several specialized components that collectively support intelligent, accurate, and adaptive clinical decision making.

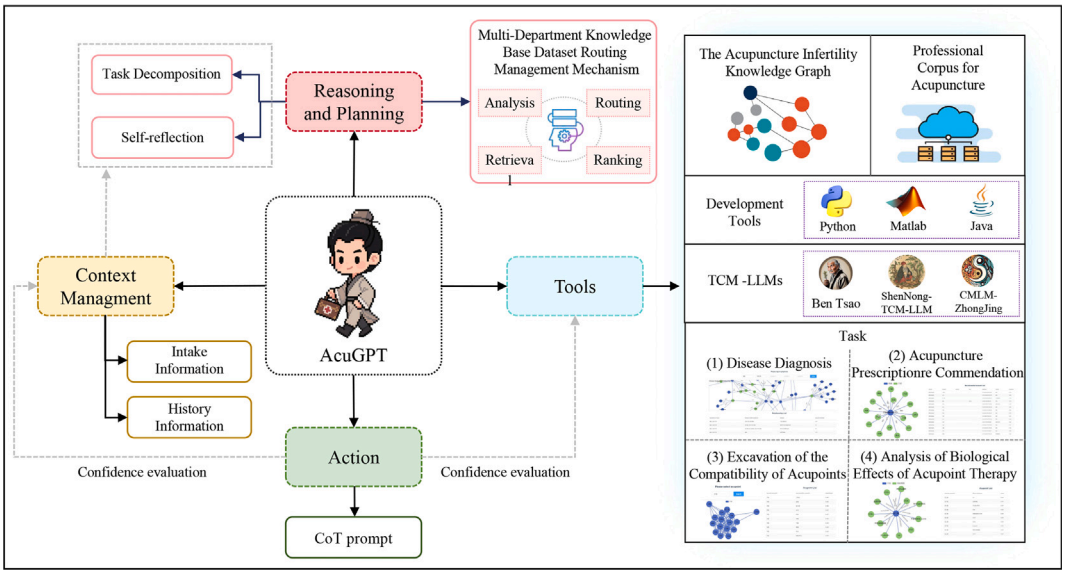

Fig. 2. Architecture of the AcuGPT-agent. A framework centered on AcuGPT for applications in acupuncture. It integrates reasoning and planning, tool utilization, context management, and action modeling to support tasks such as diagnosis, acupuncture prescription, and acupoint analysis.

The Acupuncture Infertility Knowledge Graph (ACUI-KG) is a high-quality, structured, and semantically rich knowledge graph designed to support intelligent decision making in the treatment of acupuncture-based infertility within the TCM framework. The ACUIKG defines seven categories of core medical entities: disease, symptom, syndrome, examination, needling and moxibustion method, acupoint, and effector substance, along with ten types of semantic relations, including "acts on," "stimulates," "diagnosed with," "syndrome diagnosed with", "triggers", "causes", "compatible with," "treatment method," and "associated with." Together, these components constitute a systematic ontology for acupuncture-based infertility treatment. At the data level, the construction of ACUIKG integrates diverse and heterogeneous data sources, including bilingual (Chinese and English) medical literature on acupuncture for infertility, prospective clinical observation data, clinical guidelines, and standardized textbook content. The extraction of entities and relation quintuplets was performed using GPT-4o in conjunction with expert-curated prompts developed by domain specialists. The quality and effectiveness of the extraction strategy were rigorously validated through expert evaluation. Through ontology-driven data normalization, relation property enrichment, and cross-lingual entity alignment, the resulting ACUIKG comprises 8,516 entity nodes and 32,139 semantic relations. As a key knowledge resource within the Agent system, ACUIKG enables fine-grained modeling and semantic integration of acupuncture treatment processes while providing a robust foundation for downstream tasks such as intelligent question answering, personalized recommendation, and diagnostic assistance.

Professional Corpus for Acupuncture: A large collection of professional literature, ancient texts, clinical cases, and other materials related to acupuncture is compiled to provide the agent with a rich source of knowledge, allowing him to make reasoning and judgments based on authoritative knowledge.

Development Tools: Use programming languages and development tools such as Python, Matlab, and Java to provide computational capabilities, data analysis capabilities, data visualization, and other capabilities for the agent to assist patients.

TCM - LLMs: The tools module can leverage models such as Ben Tao, ShenNong - TCM - LLM, and CMLM - Zhongjing, which are equipped with extensive knowledge of TCM and strong language understanding capabilities, to enhance the professionalism and accuracy of AcuGPT-Agent.

Tasks: Define the specific tasks related to acupuncture that the agent must perform, including diagnosis of the disease, recommendations for the prescription of acupuncture, exploration of combination acupuncture points, analysis of biological effects of points acupuncture, and others. These tasks cover the entire process of acupuncture treatment and its mechanisms. Many key pieces of information in the field of acupuncture (such as potential correlations between symptoms and diseases and synergistic mechanisms of acupuncture point combinations) are difficult to obtain directly through logical reasoning and require reliance on professional corpora and deep intelligent modeling. AcuGPT-Agent enhances the system's intelligent processing capabilities in diagnosis and treatment assistance, solution optimization, and knowledge discovery through modular design and data-driven methods. The computational methods used for each subtask include the following:

(1) Disease Diagnosis Algorithm Based on Set Similarity

In infertility diagnosis, AcuGPT-Agent assists clinicians in identifying potential diseases by calculating the Jaccard similarity between the patient's input symptoms and the symptom sets associated with known diseases.The calculation formula is as follows:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{8}$$

where, A is the set of symptoms input by the patient, B is the set of symptom diagnoses corresponding to the disease to be matched in the database, $|A \cap B|$ represents the intersection size of the input symptoms and the set of symptoms of the disease to be matched, and $|A \cup B|$ represents the union size of the two.

By calculating the similarity of Jaccard, the agent quickly identifies the diseases that are most closely aligned with the patient's symptoms, providing clinicians with the most likely diagnosis and recommending personalized acupuncture treatment plans accordingly.

(2) Acupuncture Prescription Recommendation Algorithm Based on Network Topology Analysis and Dynamic Association Rule Matching

To enable personalized acupuncture prescription recommendations, the AcuGPT-Agent decomposes the task into two key subtasks, supported by network topology analysis and association rule algorithms, and implements a two-stage recommendation mechanism for core acupoint identification and stimulation-scheme matching.

Stage One involves performing a pagerank computation on the "disease–acupoint" subgraph to select the top 15 ranked acupoints. The

formula is as follows:

$$PR(a) = (1-d) + d \sum_{p \in \mathrm{In(a)}} \frac{PR(p)}{C(p)}, \tag{9}$$

where, $PR(a)$ denotes the pagerank value of acupoint a, d is the damping factor (typically 0.85), $\mathrm{In}(a)$ represents all disease nodes pointing to acupoint a, $PR(p)$ is the pagerank value of disease p pointing to a, and $C(p)$ is the number of acupoints pointed to by disease p.

In Stage Two, the system uses association rule analysis to identify the most common needling and moxibustion methods associated with core acupoints. By mining frequent itemsets of acupoints and treatment methods, it selects the highest confidence combinations for each acupoint. The confidence score is calculated as follows:

$$\mathrm{Confidence}(a \rightarrow m) = \frac{\mathrm{support}(a \cap m)}{\mathrm{support}(a)}, \tag{10}$$

where, $\mathrm{support}(a)$ denotes the frequency of acupoint a in the dataset, $\mathrm{support}(a \cap m)$ indicates the frequency of co-occurrence of acupoint a and the needling and moxibustion method m. Through this comprehensive two-stage analysis, the system offers scientifically grounded, personalized, and clinically relevant acupuncture prescriptions for infertility patients.

(3) Acupoint Compatibility Weight Evaluation Algorithm

Based on ACUIKG, the AcuGPT-Agent analyzes synergistic effects between acupuncture points to identify high-frequency compatibility combinations for co-occurrence. The compatibility weight is calculated using the following formula:

$$W_{ah} = \frac{f_{ah}}{\sum_k f_{ak}}, \tag{11}$$

where, W_{ah} represents the compatibility weight between acupoints a and h, f_{ah} denotes the number of co-occurrences between acupoints, $\Sigma_k f_{ak}$ represents the total frequency of acupoint a with all other acupoints. A higher value indicates a more frequent clinical co-occurrence and a higher reference value. Using these weights, the system ranks acupoint combinations by priority and provides optimized compatibility schemes.

(4) Quantitative Algorithm for the Biological Effects of Acupoints.

Based on ACUIKG, the AcuGPT-Agent quantitatively models the biological effects of acupuncture points by analyzing the stimulus relationships between acupoints and effector substances. It uses TCM theory and the modern biomedical literature to identify and evaluate associations between acupoints and therapeutic outcomes within the knowledge graph. The effect strength of acupoint a on effector substance x is computed as follows:

$$E_{ax} = \frac{f_{ax}}{\sum_y f_{ay}}, \tag{12}$$

where, E_{ax} reflects the effect strength of the acupoint a in the regulation of effector substance x, f_{ax} represents the frequency with which acupoint a stimulates effector substance x, $\Sigma_y f_{xy}$ denotes the total number of times acupoint a stimulates all effector substances across studies. Higher values indicate a more significant influence and reveal its underlying therapeutic mechanisms, which support personalized precision treatment.

3.5.2. Reasoning and planning module

The reasoning and planning module constitutes a central component of AcuGPT-Agent, designed to address the complexity and domain specificity of diagnostic and therapeutic workflows. It integrates the following three submodules:

Task Decomposition: The system uses a hierarchical decomposition strategy to manage the multistep structure of acupuncture tasks like diagnosis and prescription. Complex tasks are broken into subtasks such as disease identification, acupoint selection, and technique matching, enabling modular processing, enhancing clarity, and supporting targeted optimization.

Self-Reflection: The system incorporates an automated self-reflection mechanism that evaluates both the logical integrity and clinical validity of the output. When discrepancies are detected, it triggers re-execution or additional retrieval, facilitating iterative refinement and aligning the system's reasoning process with that of human experts.

MDKBRMM: Given that medical knowledge is inherently compartmentalized across distinct specialties, we propose an LLM-powered routing mechanism to ensure precise alignment between user intent and the corresponding domain-specific expertise. This approach enables the system to accurately interpret user queries, determine the most relevant medical department, and retrieve specialized knowledge, thus improving both the relevance and precision of the system's response.

The proposed mechanism operates through a three-stage pipeline. In the first stage, knowledge base construction, the system utilizes the DeepSeekV2 model to generate a structured medical question-answer dataset from the knowledge base to be retrieved, in accordance with the requirements of RAG. This dataset is then processed in batches and semantically categorized using a foundation model (e.g., Qwen2-7B-Instruct), allowing the construction of specialized knowledge bases corresponding to various medical departments. In the second stage, intent analysis and routing, the system analyzes the user query to determine whether the available input is contextually sufficient for reliable decision making. If the input is incomplete, the system autonomously acquires additional contextual information. Once deemed adequate, the query is semantically interpreted and routed to the most relevant department-level knowledge base. In the third stage, retrieval and ranking, a fine-tuned dense retriever model (bge-large-zh-v1.5) first retrieves a set of candidate knowledge entries, then concatenates each candidate with the query, and feeds the pair into a fine-tuned BERT cross-encoder to compute relevance scores and reorder the entries in descending order, ultimately selecting the top five. This dual-model architecture balances retrieval efficiency and precision, enabling the system to accurately pinpoint knowledge within a vast and heterogeneous knowledge space.

To illustrate this pipeline in practice, we present the following input: "A 32-year-old female with a history of miscarriage. She has delayed menstruation with a moderate flow of light color and thin texture,…, What underlying causes or syndromes might the patient's symptoms suggest?" Upon receiving this query, the system performs intent recognition and analysis. The query is then classified and routed to the most relevant department-level knowledge bases, including reproductive endocrinology, gastroenterology, gynecology of TCM, and neurology. A dense retriever and BERT-based cross-encoder are subsequently used to rank candidate entries, and the top five results are returned, as illustrated in Fig. 3.

The application of this mechanism in this study is supported by a carefully curated 1GB multi-department corpus of high-quality doctor-patient interactions, which serves as the retrievable knowledge base, enabling scalable and context-aware reasoning across diverse medical specialties.

3.5.3. Action module

The action module handles the agent's decision-making process and is guided by Chain of Thought (CoT) prompting to take concrete action steps to achieve specific task objectives. In the context of infertility treatment tasks, the AcuGPT-Agent follows a structured multistage process.

The first step is to collect as much information from the patient as possible, such as medical history, including menstrual irregularities, previous treatments, and quantifiable lifestyle factors such as age, weight, blood sugar, and blood pressure, and then to evaluate the confidence of the model, reflecting on whether the patient's input exceeds the knowledge base of the model. If it passes, AcuGPT uses its knowledge of TCM

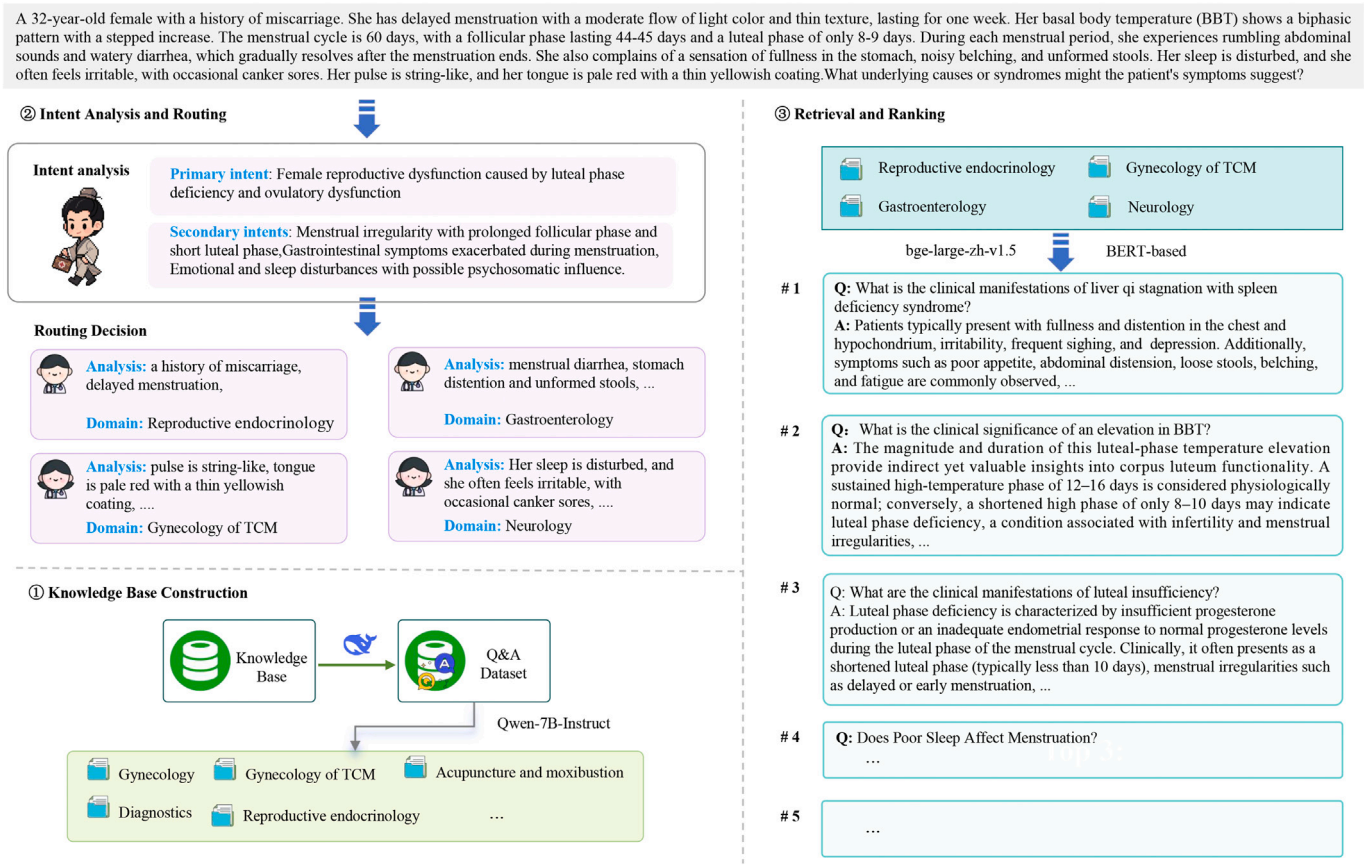

Fig. 3. Illustrative example of our proposed multi-department knowledge base routing management mechanism (MDKBRMM) framework.

acupuncture to respond to the patient, for example, by explaining professional terminology or engaging in general communication. If evaluation fails, the second step is triggered. The second step involves calling the context management module, which reintroduces all communication from the first step to AcuGPT. The model then reflects and reasons, generating the appropriate tool invocation methods and their sequence. The third step involves calling the Tools module to further supplement the knowledge and generate a response, addressing the patient's needs.

4. Experiments

4.1. Performance evaluation

In the current research landscape, the field of TCM acupuncture lacks a standardized and authoritative evaluation benchmark. To systematically assess the domain expertise and practical application capabilities of our AcuGPT model, we constructed a specialized evaluation dataset named EvalAcu. This dataset enables a comprehensive evaluation of domain-specific competencies.

We summarize an algorithm, illustrated in Fig. 4 and encapsulated as Algorithm 1, to construct the EvalAcu dataset. The process consists of several interconnected stages. First, we collected texts from TCM university exams, professional certification tests, and curated clinical Q&A by acupuncture practitioners to form a foundational corpus. Then, we utilized the DeepSeek-V2 language model to segment the corpus and extract essential knowledge elements. These key knowledge were transformed into vector representations using the BGE-M3 embedding model, followed by subset selection using the KCenterGreedy algorithm. Then DeepSeek-V2 was used again to automatically generate single-choice

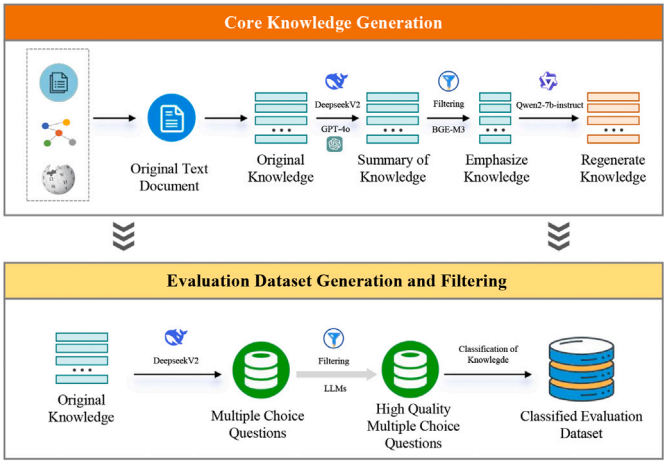

Fig. 4. The workflow for constructing the EvalAcu dataset.

questions from the selected knowledge. Multiple language models, including Qwen2-7B-Instruct and GLM-4-9B-Chat, were used to test the difficulty of the questions, retaining those frequently answered incorrectly. Finally, GPT-4o rigorously assessed each retained question in terms of factual accuracy, logical consistency, and truthfulness. The lowest scoring 20 % of the questions were eliminated and the remaining high-quality questions were classified by knowledge type, completing the construction of the EvalAcu dataset.

Algorithm 1 Construction of EvalAcu dataset.

Require: Raw corpus C from domain-related literature
Ensure: Curated evaluation dataset $\mathcal{D}_{EvalAcu}$

Corpus Collection: Crawl and aggregate domain-specific text corpus C from academic literature.
Knowledge Extraction:
for each document $d \in C$ **do**
    Apply LLM extractor $f_{extract}$ (e.g., DeepSeek-V2) to obtain knowledge $K = \{k_1, \ldots, k_n\}$.
end for
Embedding & Subset Selection:
Compute embeddings $v_k = f_{vec}(k)$ for all $k \in K$ using BGE-M3.
Apply KCenterGreedy to select representative subset $K' \subset K$.
Question Generation:
for each $k_i \in K'$ **do**
    Generate multiple-choice $q_i = f_{gen}(k_i)$ using an instruction-tuned LLM.
end for
Filtering by LLM Committee:
Evaluate each q_i with LLM $\mathcal{M} = \{m_1, \ldots, m_k\}$.
Retain q_i if at least τ models answer incorrectly.
Quality Scoring and Pruning:
Use GPT-4o to evaluate questions on: factual accuracy, logical consistency, and truthfulness.
Remove bottom 20 % based on aggregate score.
Final Classification:
Classify remaining questions into *Simple* and *Difficult* subsets based on zero-shot LLM accuracy.
return $\mathcal{D}_{EvalAcu} = \mathcal{D}_{simple} \cup \mathcal{D}_{difficult}$

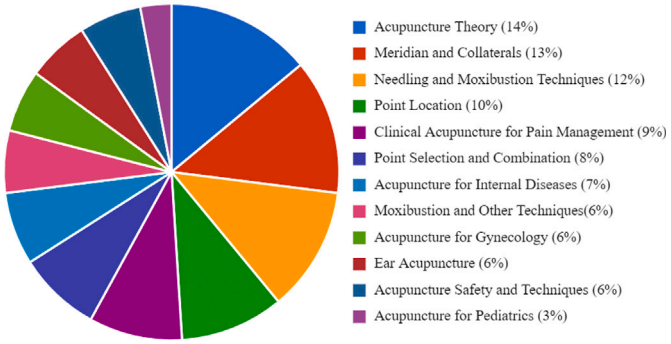

Fig. 5. Composition of the EvalAcu dataset. The EvalAcu dataset is a domain-specific benchmark for assessing language model competencies across 12 subfields of acupuncture.

The EvalAcu dataset (Fig. 5) covers 12 subfields of acupuncture and contains 10,202 multiple-choice questions. To analyze the performance of the model during professional knowledge learning, we divided it into two subsets. The simple dataset, which contains 1,260 questions that all evaluated models can answer correctly in zero-shot mode, serves as a baseline to measure the model's grasp of fundamental acupuncture knowledge. The difficult dataset, with 1,830 questions, consists of questions that no evaluated model can answer correctly directly in zero-shot mode, and elimination of the lowest score 20 % of the questions. As shown in Table 3, after three rounds of rapid filtering, a difficult high quality evaluation dataset is finally constructed.

Evaluating the performance of the trained LLM is crucial for model development. For AcuGPT, we assess both its effectiveness after integrating professional acupuncture knowledge and its general capabilities across various tasks. In terms of general capability evaluation, we adopted C-Eval [51], a comprehensive evaluation suite for Chinese foundation models, to measure the model's language understanding and reasoning abilities in non-acupuncture tasks. Regarding medical general capability evaluation, we used three benchmark datasets: MEDQA [52], a USMLE-based multiple choice question dataset to assess clinical reasoning and medical knowledge; MEDMCQA [53], derived from Indian medical exams that cover basic medicine, physiology and pathology; and PUBMEDQA [54], constructed from the PubMed literature to assess the model's ability to understand and answer questions based on real-world biomedical texts.

Table 3
Comparison of models' accuracy.

Model	Access	Round 1	Round 2	Round 3
GPT-4o	API	75.66	49.27	50.90
DeepSeek-V2	API	79.13	56.74	58.39
ChatGLM-4	API	77.87	53.98	55.80
ChatGLM4-9B-Chat	Weights	73.51	45.53	47.44
InternLM2.5-7B-Chat	Weights	71.95	41.89	43.91
Qwen2-7B-Instruct	Weights	75.56	49.37	51.08

Table 4
Evaluation of general and domain-specific capabilities of different LLMs.

Train type	Train dataset	Model	EvalAcu	C-Eval
/	/	DeepSeek-V2	55.17	56.32
/	/	ChatGLM3-6B-32 k	55.13	59.14
/	/	InternLM2-Chat-7B	59.15	60.61
/	/	Qwen2-7B-Chat	63.15	60.23
/	/	Qwen2-7B-Instruct	65.23	**70.27**
Lora	SFT dataset	DeepSeek-V2	61.16	52.16
Lora	SFT dataset	ChatGLM3-6B-32 k	65.19	56.22
Lora	SFT dataset	InternLM2-Chat-7B	68.13	57.27
Lora	SFT dataset	Qwen2-7B-Chat	68.46	56.21
Lora	SFT dataset	Qwen2-7B-Instruct	**70.15**	54.25

4.2. Base models

In our experiments, we selected five open-source models as base models for training, namely Qwen2-7B-Chat, Qwen2-7B-Instruct, InternLM2-Chat-7B, ChatGLM3-6B-32 K, and DeepSeek-V2. To evaluate both general and domain-specific capabilities, we performed comparative evaluations using the C-Eval benchmark and the domain-specific EvalAcu dataset. As shown in Table 4, the evaluation results indicate that, among the benchmark models, Qwen2-7B-Instruct performed the best, followed by Qwen2-7B-Chat and InternLM2-Chat-7B, while DeepSeek V2 and ChatGLM3-6B performed relatively poorly. In the initial experimental phase, all models were fine-tuned on the same SFT dataset using the LoRA method. Notably, Qwen2-7B-Instruct and InternLM2-Chat-7B showed substantial improvements, reaching accuracy scores close to 70 %. In contrast, ChatGLM3-6B reached only 65 % accuracy, reflecting more limited learning capacity under the same conditions. These results highlight Qwen2-7B-Instruct as the most effective model for domain-specific knowledge acquisition in the acupuncture field. Therefore, it was selected as the benchmark base model for subsequent comparative experiments and further optimization.

We performed SFT training on two A800 GPUs, with all local model inference conducted on a single A800 GPU. This training used our custom-built Acupuncture SFT dataset, tailored specifically to acupuncture-related content. During the SFT training process, we configured the learning rate, batch size, maximum context length, and other hyperparameters, including LoRA, as detailed in Table 5. To further bolster general domain proficiency, we incorporated a general supplementary dataset comprising 500,000 entries and proceeded with mixed-dataset SFT training, followed by an evaluation of its efficacy.

Table 5
Training parameters.

Hyperparameter	Value
Precision	fp16
Epochs	10
Maximum context length	1024
Batch size	16
Lora rank	8
Lora alpha (scaling factor)	16
Optimizer	Adamw_torch
Learning rate regulator	cosine
Learning rate	5×10^{-5}

4.3. Training results

During the initial phase, all models were fine-tuned using the LoRA method. Among them, Qwen2-7B-Instruct exhibited superior performance in domain-specific learning and was consequently selected as the benchmark model for further comparative analysis.

To further explore optimization strategies, we conducted incremental pre-training and comprehensive fine-tuning experiments using Qwen2-7B-Instruct. The hyperparameters used for DoRA were consistent with those for LoRA, and detailed comparison results are summarized in Table 6. For acupuncture tasks, models trained with the DoRA and LoRA + methods consistently outperformed those that used LoRA alone. However, while LoRA+ fine-tuning boosted domain-specific performance, it significantly degraded results on the C-Eval general language benchmark, indicating a trade-off between specialization and generalization. In contrast, the DoRA-trained model not only excelled in the acupuncture domain, but also maintained robust performance on general benchmarks. The Qwen2-7B-Instruct model on a mixed dataset, which included a 1:5 ratio of SFT to general additional data, did not produce significant improvements in acupuncture domain evaluation or overall performance. In fact, the results were even inferior to training solely on the SFT dataset. Ultimately, the optimal balance between domain expertise and general capabilities was achieved through a three-cycle pre-training process followed by LoRA fine-tuning on a mixed dataset (a 1:1 ratio of SFT to general supplementary data). These findings guided the selection of the Pretrain + SFT strategy (a 1:1 mix-data ratio) for the final fine-tuning phase, resulting in the development of the AcuGPT model.

4.4. Evaluation of AcuGPT-Agent

AcuGPT-Agent, using prompt engineering and a modular architecture, integrates natural language input, semantic parsing, task recognition, knowledge matching, and AcuGPT invocation. Taking infertility-related queries as an example, it achieves an automated processing workflow from the main complaint analysis to the personalized recommendation of acupuncture treatment, as illustrated in Fig. 6. The user logs into the system and inputs the following inquiry: “A 35-year-old female has been married for three years without using contraception and has not been able to conceive. Over the past six months, her menstrual cycle has shortened, and the menstrual flow has decreased. She also reports experiencing hot flashes and night sweats, along with emotional symptoms such as depression, poor mental state, and persistent anxiety, particularly at night, accompanied by difficulty falling asleep. What underlying causes might these symptoms suggest and how can acupuncture be used as a therapeutic intervention?” Upon receiving this input, AcuGPT-Agent first performs semantic analysis to extract key information. Then it checks for historical data or long-term memory associations. Based on this, the agent initiates its reasoning and reflection mechanism. Through task decomposition and self-assessment, it evaluates whether external tools need to be called, confirms available resources and adaptable modules, and proceeds to the formal task execution process.

4.4.1. Domain capability evaluation

To verify that the proposed agent system can understand the problems and make the necessary decisions to invoke the appropriate modules of the tool to address infertility issues, we selected 120 infertility related questions from EvalAcu to create EvalInfer and subsequently conducted systematic comparative tests. Based on the results in Table 7, it is evident that AcuGPT and AcuGPT-Agent demonstrate stronger domain adaptability in specialized acupuncture-related question–answering scenarios compared to other large medical models.

We selected several high-performing LLMs commonly used in medical question answering, such as HuaTuoGPT-II [55], Meditron-70B [56], Med-PaLM [57], Gemini Pro [58], as baselines and compared them with our proposed models, AcuGPT and AcuGPT-Agent. The results demonstrate that AcuGPT-Agent exhibits the most balanced and outstanding performance across all datasets, achieving the highest scores in four out of five tasks: MEDQA (69.39 %), MEDMCQA (67.82 %), EvalAcu (81.32 %), and EvalInfer (80.00 %). These results clearly surpass all the baseline methods. This indicates that AcuGPT-Agent possesses excellent generalization capabilities and architectural robustness across both domain-specific and general medical question answering tasks.

Meanwhile, although AcuGPT performs slightly below Med-PaLM and Gemini Pro in general medical datasets such as MEDQA (54.39 %) and MEDMCQA (56.82 %), it outperforms all models in acupuncture-specific tasks, achieving the highest scores on EvalAcu (78.13 %) and EvalInfer (55.83 %). This shows that our fine-tuned model and the designed agent architecture significantly enhance domain-specific competence in acupuncture, while also maintaining competitive performance in general medical tasks. Together, they offer a comprehensive solution for achieving expert-level reasoning in acupuncture knowledge as well as broad medical question answering.

4.4.2. Model performance

To construct a more rigorous evaluation framework, we used GPT-4o to filter 2943 open-ended questions from the EvalAcu database. These questions required free-form responses from the models, making them suitable for evaluating subjective question answering. We invited three Ph.D. students from Nanjing University of Chinese Medicine to evaluate the responses generated by relatively better performing open source medical models, Med-PALM, AcuGPT and AcuGPT-Agent, as shown in Fig. 7.

Table 6
Comparison of different SFT methods.

Train type	Train dataset	Model	EvalAcu	C-Eval
Pretrain (LoRA) + SFT (LoRA)	Pre-training dataset + SFT dataset	Qwen2-7B-Instruct	71.26	54.26
Pretrain (LoRA) + SFT (LoRA)	Pre-training dataset + Mix-data (1:1)	Qwen2-7B-Instruct	73.31	54.21
Pretrain (LoRA) + SFT (LoRA)	Pre-training dataset + Mix-data (1:5)	Qwen2-7B-Instruct	72.13	55.13
Pretrain + SFT	Pre-training dataset + Mix-data (1:1)	Qwen2-7B-Instruct	**78.13**	**62.15**
LoRA	SFT dataset	Qwen2-7B-Instruct	72.13	53.14
DoRA	SFT dataset	Qwen2-7B-Instruct	75.73	56.17
LoRA+	SFT dataset	Qwen2-7B-Instruct	73.31	50.78

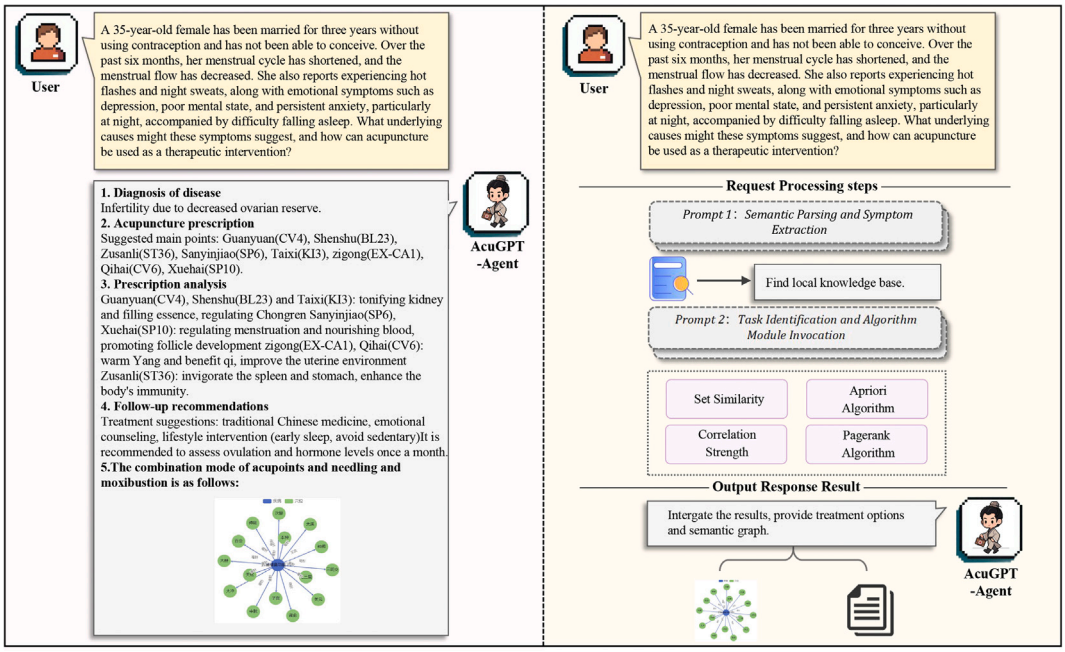

Fig. 6. AcuGPT-Agent task processing. The left panel illustrates the interaction between the user and AcuGPT-Agent during a medical consultation. The right panel outlines the computational steps involved in AcuGPT-Agent's response generation.

Table 7
Domain capability evaluation of different methods.

Method	MEDQA	MEDMCQA	PUBMEDQA	EvalAcu	EvalInfer
HuaTuoGPT-II	41.13	41.87	78.32	52.86	40.24
Meditron	58.53	48.28	76.78	55.36	42.69
Med-PALM	67.61	57.74	**79.44**	53.20	46.72
Gemini-Pro	58.23	54.54	71.39	50.89	38.63
AcuGPT	54.39	56.82	73.21	78.13	55.83
AcuGPT-Agent	**69.39**	**67.82**	75.21	**81.32**	**80.00**

To ensure consistency and fairness in the scoring, the evaluation was performed in five key dimensions: precision, relevance, logical consistency, expertise, and fluency. Accuracy refers to whether the response is factually correct and aligned with established medical knowledge, particularly in the context of acupuncture and infertility treatment. Relevance evaluates how well the response addresses the core of the question and whether it provides a complete answer. Logical consistency assesses the coherence of reasoning, the clarity of thought, and the internal logic within the response. Expertise measures the depth of domain-specific knowledge demonstrated in the answer, with a focus on professional acupuncture practices and clinical treatment for infertility. Fluency considers whether the language is smooth, grammatically correct, and easy to understand. Each dimension is rated on a scale of 1 to 5, where 1 indicates very poor performance and 5 represents excellent performance. The total score for each response is the sum of these five dimensions, with a maximum possible score of 25 points.

Based on the scoring results in Fig. 7, Med-PALM performs poorly in accuracy and relevance. AcuGPT shows a significant improvement in the domain-specific area, especially in relevance, where it scores 5, indicating that its responses are more aligned with user needs in domain-related questions. However, AcuGPT scores slightly lower in logical consistency and expertise. AcuGPT-Agent, on the other hand, outperforms both Med-PALM and AcuGPT in accuracy, relevance, logical consistency, expertise, and fluency, ultimately achieving a score of 23. This shows its strong advantage in specialized domains.

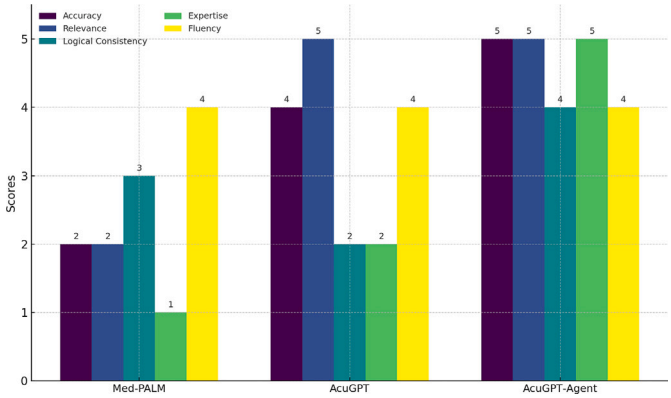

Fig. 7. Performance comparison of different models.

4.4.3. Ablation study

To further assess the contribution of the MDKBRMM mechanism to the retrieval quality and task accuracy, we performed an ablation study by comparing the full model with a variant in which the MDKBRMM module was removed. The comparative results across datasets are reported in Table 8.

The data indicate that the removal of the MDKBRMM module leads to a substantial decrease in performance across all evaluated datasets. Specifically, the ablated model AcuGPT-Agent (w/o

Table 8
Ablation study on the MDKBRMM module.

Method	MEDQA	MEDMCQA	PUBMEDQA	EvalAcu
AcuGPT-Agent (w/o MDKBRMM)	46.72	50.69	68.71	72.34
AcuGPT-Agent (Full)	**69.39**	**67.82**	**75.21**	**81.32**

MDKBRMM) achieves accuracy scores of 46. 72 %, 50. 69 %, 68. 71 %, and 72. 34 % in MEDQA, MEDMCQA, PUBMEDQA and EvalAcu, respectively. In contrast, the complete AcuGPT-Agent system achieves 69. 39 %, 67. 82 %, 75. 21 %, and 81. 32 %. These results confirm the effectiveness of MDKBRMM in improving retrieval relevance and mitigating semantic interference in multidisciplinary medical corpora.

5. Discussion

Our results highlight the strengths of AcuGPT and AcuGPT-Agent, and demonstrate how LLMs can be adapted to support domain-specific applications in medicine. AcuGPT is developed to address the particular challenges of TCM, especially in the context of acupuncture, where clinical reasoning relies on specialized terminology and complex theoretical constructs. By applying techniques such as LoRA, DoRA, and a Pretrain + SFT pipeline using mixed-domain data, AcuGPT is able to achieve a balance between general language understanding and acupuncture-specific knowledge.

AcuGPT-Agent extends the capabilities of AcuGPT by incorporating it into a modular framework that integrates domain-specific resources, such as acupuncture knowledge graphs and MDKBRMM. This structure allows the system to route queries across specialized knowledge bases, enabling more accurate and contextually relevant responses. The use of external resources complements the LLM's ability to generate responses, ensuring that it can produce clinically relevant outputs aligned with acupuncturists' therapeutic reasoning. Compared to general-purpose or single-domain models such as Med-PALM and HuatuoGPT-II, our approach shows more consistent alignment with acupuncturists' reasoning patterns in benchmark tasks.

However, several important limitations must be addressed before this system can be deployed in real-world clinical settings. First, hallucination remains an inherent limitation of current LLMs. Although our model performs well on acupuncture-related tasks, robust safety mechanisms are critical. In future practical applications, we will explore and develop two strategies: (1) Tiered Human-AI Collaboration: Based on patient input, AI responses to common or popular science questions can be delivered directly. For rare symptoms or prescription-related inquiries, an expert review is required prior to response delivery. (2) Dynamic Confidence Evaluation: Based on model outputs, confidence is estimated using adversarial robustness, output entropy, and heuristic calibration. Responses above a predefined threshold may be returned directly; otherwise, expert validation is required.

Second, LLM evaluation benchmarks have proliferated, yet mainstream leaderboards still revolve around offline benchmark scores or Elo human preference ratings, lacking unified metrics for dimensions such as explainability and long-term stability. Relying on a handful of benchmarks cannot fully or impartially capture the true capabilities of a model, especially in specialized domains such as medicine and law. Our proposed method for automatically constructing discipline-specific test sets therefore has limitations. In the future, we will explore a more comprehensive evaluation framework that integrates patient feedback, expert review, long-term deployment performance, as well as adversarial and dynamically updated benchmarks. The goal is to develop a cross-disciplinary, robust and resource-efficient evaluation system to improve the quality of assessment under constrained conditions.

In general, our findings demonstrate that LLMs can play a constructive role in specialized medical domains by effectively interacting with domain-specific knowledge systems. Through thoughtful integration into domain workflows, LLMs support more accurate and context-sensitive decision making.

6. Conclusion

Our results highlight the practical potential of LLMs in supporting clinical decision making for acupuncture. This study represents an early step toward adapting LLMs to the linguistic and reasoning complexities inherent in TCM acupuncture. Looking to the future, this research will advance through multiple strategic enhancements. To deepen its clinical relevance, we plan to integrate multimodal patient data, such as sensor output and medical imaging, to enrich diagnostic input. We will also develop reasoning frameworks aligned with TCM epistemology to ensure consistency with traditional diagnostic logic. We hope these efforts encourage continued collaboration across disciplines to responsibly translate our findings into improved healthcare outcomes [59,60].

CRediT authorship contribution statement

Jing Wen: Data curation, Validation, Conceptualization, Software, Formal analysis, Writing – original draft, Visualization. **Diandong Liu:** Methodology, Data curation, Validation, Investigation, Writing – review & editing. **Yuxin Xie:** Methodology, Supervision, Validation, Writing – review & editing. **Yi Ren:** Software, Resources, Visualization. **Jiacun Wang:** Software, Resources, Visualization, Writing – review & editing. **Youbing Xia:** Conceptualization, Project administration, Supervision, Supervision, Validation, Writing – review & editing. **Peng Zhu:** Supervision, Funding acquisition, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Peng Zhu reports financial support was provided by the National Natural Science Foundation of China. Peng Zhu reports financial support was provided by the Social Science Foundation of Jiangsu Province. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant Nos. 72174087, 72474103), the Social Science Foundation of Jiangsu Province (Grant No. 22TQB004), the Key R&D Plan of Jiangsu Province (Grant No. BE2022712), the Nanjing University of Science and Technology Undergraduate Education Reform and Construction Project (Grant No. 2025BKJG056) and Graduate Education Teaching Reform Project (Grant No. KT2024-C10), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX24_2199).

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.neucom.2025.131116.

Data availability

Data will be made available upon request.

References

[1] F. Ifrim Chen, A.D. Antochi, A.G. Barbilian, Acupuncture and the retrospect of its modern research, Rom. J. Morphol. Embryol. 60 (2) (2019) 411–418.
[2] J.-S. Han, Y.-S. Ho, Global trends and performances of acupuncture research, Neurosci. Biobehav. Rev. 35 (3) (2011) 680–687.
[3] N. Ishizaki, T. Yano, K. Kawakita, Public status and prevalence of acupuncture in Japan, Evid. Based Complement Alternat Med. 7 (4) (2010) 493–500.
[4] A. Burke, D.M. Upchurch, C. Dye, L. Chyu, Acupuncture use in the United States: findings from the National health interview survey, J. Altern Complement. Med. 12 (7) (2006) 639–648.
[5] X. Zhong, X. Zeng, L. Zhao, C. Tao, X. Min, R. He, Clinicians' knowledge and understanding regarding multidisciplinary treatment implementation: a study in municipal public class III grade a hospitals in Southwest China, BMC Med. Educ. 23 (1) (2023) 916.
[6] W. Yu, Y. Lee, The current status of acupuncture Education and clinical practices in Taiwan, Medical Acupuncture 37 (1) (2025) 54–58.
[7] X. Ye, Y. Ren, Y. Chen, J. Chen, X. Tang, Z. Zhang, A "4D" systemic view on meridian essence: substantial, functional, chronological and cultural attributes, J. Integr. Med. 20 (2) (2022) 96–103.

[8] R. Han, J. Hu, Acupuncture: an overview on its functions, meridian pathways and molecular mechanisms, Am. J. Chin. Med. 52 (5) (2024) 1215–1244.

[9] H. Long, Y. Zhu, L. Jia, B. Gao, J. Liu, L. Liu, H. Herre, An ontological framework for the formalization, organization and usage of TCM-knowledge, BMC Med. Inform. Decis. Mak. 19 (2) (2019) 53.

[10] A. Givati, S. Berlinsky, The "disenchantment" of traditional acupuncturists in higher education, Health (London) 27 (1) (2023) 20–40.

[11] S. Li, W. Tan, C. Zhang, J. Li, H. Ren, Y. Guo, J. Jia, Y. Liu, X. Pan, J. Guo, W. Meng, Z. He, Taming large language models to implement diagnosis and evaluating the generation of LLMs at the semantic similarity level in acupuncture and moxibustion, Expert Syst. Appl. 264 (2025) 125920.

[12] Y. Li, X. Peng, J. Li, X. Zuo, S. Peng, D. Pei, C. Tao, H. Xu, N. Hong, Relation extraction using large language models: a case study on acupuncture point locations, J. Am. Med. Inform. Assoc. 31 (11) (2024) 2622–2631.

[13] Y. Ren, X. Luo, Y. Wang, H. Li, H. Zhang, Z. Li, H. Lai, X. Li, L. Ge, J. Estill, et al, Large language models in traditional Chinese medicine: a scoping review, J. Evid. Based Med. 18 (1) (2025) e12658.

[14] Y. Duan, Q. Zhou, Y. Li, C. Qin, Z. Wang, H. Kan, J. Hu, Research on a traditional Chinese medicine case-based question-answering system integrating large language models and knowledge graphs, Front Med. (Lausanne) 11 (2024) 1512329.

[15] G. Yang, X. Liu, J. Shi, Z. Wang, G. Wang, TCM-GPT: efficient pre-training of large language models for domain adaptation in traditional Chinese medicine, Comput. Methods Programs Biomed. Update 6 (2024) 100158.

[16] Y. Ren, X. Luo, Y. Wang, H. Li, H. Zhang, Z. Li, H. Lai, X. Li, L. Ge, J. Estill, L. Zhang, S. Yang, Y. Chen, C. Wen, Z. Bian, Advanced Working Group, Large language models in traditional Chinese medicine: a scoping review, J. Evid. Based Med. 18 (1, Mar. 2025) e12658.

[17] L. Zhu, W. Mou, Y. Lai, J. Lin, P. Luo, Language and cultural bias in AI: comparing the performance of large language models developed in different countries on traditional Chinese medicine highlights the need for localized models, J. Transl. Med. 22 (1) (2024) 319.

[18] M. Zarfati, S. Soffer, G.N. Nadkarni, E. Klang, Retrieval-augmented generation: advancing personalized care and research in oncology, Eur. J. Cancer 220 (2025) 115341.

[19] S.V. Shah, Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical Records, JAMA Netw. Open 7 (8) (2024) e2425953.

[20] A. Soroush, B.S. Glicksberg, E. Zimlichman, Y. Barash, R. Freeman, A.W. Charney, G.N. Nadkarni, E. Klang, Large language models are poor medical coders — benchmarking of medical code querying, NEJM AI 1 (5) (2024) AIdbp2300040.

[21] W. Li, X. Ge, S. Liu, L. Xu, X. Zhai, L. Yu, Opportunities and challenges of traditional Chinese medicine doctors in the era of artificial intelligence, Front. Med. 10 (2024).

[22] B.J. Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, Y. Su, Thinking about GPT-3 in-context learning for biomedical IE? think again, in: Findings of the Association for Computational Linguistics, EMNLP 2022, 2022, pp. 4497–4512.

[23] Z. Wang, J. Jiang, Y. Zhan, B. Zhou, Y. Li, C. Zhang, B. Yu, L. Ding, H. Jin, J. Peng, X. Lin, W. Liu, Hypnos: a domain-specific large language model for anesthesiology, Neurocomputing 624 (2025) 129389.

[24] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, J. Li, G. Chen, X. Wu, Z. Zhang, Q. Xiao, et al, HuaTuoGPT, towards taming language model to be a doctor, in: Findings of the Association for Computational Linguistics, EMNLP 2023, 2023, pp. 10859–10885.

[25] J. Chen, X. Wang, K. Ji, A. Gao, F. Jiang, S. Chen, H. Zhang, D. Song, W. Xie, C. Kong, et al, HuaTuoGPT-II, one-stage training for medical adaption of LLMs, arXiv preprint arXiv:2311.09774, 2023.

[26] Y. Tan, Z. Zhang, M. Li, F. Pan, H. Duan, Z. Huang, H. Deng, Z. Yu, C. Yang, G. Shen, P. Qi, C. Yue, Y. Liu, L. Hong, H. Yu, G. Fan, Y. Tang, MedChatZH: a tuning LLM for traditional Chinese medicine consultations, Comput. Biol. Med. 172 (2024) 108290.

[27] Y. Dai, X. Shao, J. Zhang, Y. Chen, Q. Chen, J. Liao, F. Chi, J. Zhang, X. Fan, TCMChat: a generative large language model for traditional Chinese medicine, Pharmacol. Res 210 (2024) 107530.

[28] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, 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 Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, vol. 38, 2024, pp. 19368–19376.

[29] A. Penzias, R. Azziz, K. Bendikson, M. Cedars, T. Falcone, K. Hansen, M. Hill, S. Jindal, S. Kalra, J. Mersereau, et al, Fertility evaluation of infertile women: a committee opinion, Fertil. Steril. 116 (5) (2021) 1255–1265.

[30] Y. Wei, Z. Lin, Q. Huang, H. Wu, R. Wang, J. Wang, Burden of female infertility in 204 countries and territories, 1990–2021: results from the global burden of disease study 2021, J. Psychosom. Obstet. Gynaecol. 46 (1) (2025) 2459618.

[31] S.A. Carson, A.N. Kallen, Diagnosis and management of infertility: a review, JAMA 326 (1) (2021) 65.

[32] M. Herbert, M. Choudhary, D. Zander-Fox, Assisted reproductive technologies at the nexus of fertility treatment and disease prevention, Science 380 (6641) (2023) 164–167.

[33] L. Hickstein, S. Kiel, C. Raus, S. Heß, J. Walker, J.-F. Chenot, Acupuncture covered by statutory health insurance in Germany: an observational study based on claims data, Schmerz 32 (1) (2018) 30–38.

[34] H. Li, X. Jin, P.M. Herman, C.M. Witt, Y. Chen, W. Gang, X. Jing, P. Song, L. Yang, D. Ollendorf, et al, Using economic evaluations to support acupuncture reimbursement decisions: current evidence and gaps, BMJ 376 (2022) e067477.

[35] L. Zhao, M. Sun, Z. Yin, J. Cui, R. Wang, L. Ji, G. Geng, J. Chen, D. Cai, Q. Liu, et al, Long-term effects of individualized acupuncture for chronic neck pain: a randomized controlled trial, Ann. Intern. Med. 177 (10) (2024) 1330–1338.

[36] W.E. Paulus, M. Zhang, E. Strehler, I. El-Danasouri, K. Sterzik, Influence of acupuncture on the pregnancy rate in patients who undergo assisted reproduction therapy, Fertil. Steril. 77 (4) (2002) 721–724.

[37] E. Stener-Victorin, U. Waldenström, L. Nilsson, M. Wikland, P.O. Janson, A prospective randomized study of electro-acupuncture versus alfentanil as anaesthesia during oocyte aspiration in in-vitro fertilization, Hum. Reprod. 14 (10) (1999) 2480–2484.

[38] C. Zhu, W. Xia, J. Huang, X. Zhang, F. Li, X. Yu, J. Ma, Q. Zeng, Effects of acupuncture on the pregnancy outcomes of frozen-thawed embryo transfer: a systematic review and meta-analysis, Front. Public Health 10 (2022) 987276.

[39] X. Feng, N. Zhu, S. Yang, L. Wang, W. Sun, R. Li, F. Gong, S. Han, R. Zhang, J. Han, Transcutaneous electrical acupoint stimulation improves endometrial receptivity resulting in improved IVF-ET pregnancy outcomes in older women: a multicenter, randomized, controlled clinical trial, Reprod. Biol. Endocrinol. 20 (1) (2022) 127.

[40] A.J. Vickers, E.A. Vertosick, G. Lewith, H. MacPherson, N.E. Foster, K.J. Sherman, D. Irnich, C.M. Witt, K. Linde, Acupuncture Trialists' Collaboration, Do the effects of acupuncture vary between acupuncturists? Analysis of the acupuncture trialists' collaboration individual patient data meta-analysis, Acupunct. Med. 39 (4) (2021) 309–317.

[41] S. Cochrane, C.A. Smith, A. Possamai-Inesedy, Development of a fertility acupuncture protocol: defining an acupuncture treatment protocol to support and treat women experiencing conception delays, J. Altern Complement. Med. 17 (4) (2011) 329–337.

[42] T.Y. Chon, M.C. Lee, Acupuncture, Mayo Clin. Proc. 88 (10) (2013) 1141–1146.

[43] H. Li, J.E. Darby, I. Akpotu, J.M. Schlaeger, C.L. Patil, O. Danciu, A.D. Boyd, L. Burke, M.O. Ezenwa, M.R. Knisely, et al, Barriers and facilitators to integrating acupuncture into the US health care system: a scoping review, J. Integr. Complement Med. 30 (12) (2024) 1134–1146.

[44] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: Findings of the Association for Computational Linguistics, ACL 2024, 2024, pp. 2318–2335.

[45] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, tech. rep., OpenAI, 2018.

[46] B. Hanindhito, B. Patel, L.K. John, Large language model Fine-tuning with low-rank adaptation: a performance exploration, in: Proceedings of the 16th ACM/SPEC International Conference on Performance Engineering, ICPE '25, 2025, pp. 92–104.

[47] Y. Mao, K. Huang, C. Guan, G. Bao, F. Mo, J. Xu, DoRA: enhancing parameter-efficient fine-tuning with dynamic rank distribution, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, vol. 1, 2024, pp. 11662–11675.

[48] S. Hayou, N. Ghosh, B. Yu, LoRA+: efficient low rank adaptation of large models, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, 2024.

[49] C. Xin, Y. Lu, H. Lin, S. Zhou, H. Zhu, W. Wang, Z. Liu, X. Han, L. Sun, Beyond full fine-tuning: harnessing the power of LoRA for multi-task instruction tuning, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024, 2024, pp. 2307–2317.

[50] W. Lu, R.K. Luu, M.J. Buehler, Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities, NPJ Comput. Mater 11 (1) (2025) 84.

[51] Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, Y. Fu, et al, C-Eval: a multi-level multi-discipline Chinese evaluation suite for foundation models, Adv. Neural Inf. Process. Syst. 36 (2023) 62991–63010.

[52] D. Jin, E. Pan, N. Oufattole, W. Weng, H. Fang, P. Szolovits, What disease does this patient have? A large-scale open domain question answering dataset from medical exams, arXiv preprint arXiv:2009.13081, 2020.

[53] A. Pal, L.K. Umapathi, M. Sankarasubbu, MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering, in: Proceedings of the Conference on Health, Inference, and Learning, vol. 174, PMLR, 2022, pp. 248–260.

[54] Q. Jin, B. Dhingra, Z. Liu, W.W. Cohen, X. Lu, PubMedQA: a dataset for biomedical research question answering, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 2567–2577.

[55] J. Chen, X. Wang, K. Ji, A. Gao, F. Jiang, S. Chen, H. Zhang, D. Song, W. Xie, C. Kong, et al, HuaTuoGPT-II, one-stage training for medical adaption of LLMs, arXiv preprint arXiv:2311.09774, 2023.

[56] Z. Chen, A.H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, et al, MEDITRON-70B: scaling medical pretraining for large language models, arXiv preprint arXiv:2311.16079, 2023.

[57] 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 620 (7972) (2023) 172–180.

[58] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al, Capabilities of gemini models in medicine, arXiv preprint arXiv:2404.18416, 2024.

[59] H. Li, Z. Wang, N. Zeng, P. Wu, Y. Li, Promoting objective knowledge transfer: a cascaded fuzzy system for solving dynamic multiobjective optimization problems, IEEE Trans. Fuzzy Syst. 32 (11) (2024) 6199–6213.

[60] P. Wu, H. Li, X. Luo, L. Hu, R. Yang, N. Zeng, From data analysis to intelligent maintenance: a survey on visual defect detection in aero-engines, Meas. Sci. Technol. 36 (6) (2025) 062001.

Author biography

Jing Wen is currently pursuing her Ph.D. degree in Acupuncture and Moxibustion at Nanjing University of Chinese Medicine. Her research focuses on integrating acupuncture with intelligent technologies, including the development of knowledge graphs for acupuncture-based infertility treatments and the application of large language models in the field. She has contributed to various research projects and publications in integrative medicine, with the aim of bringing together traditional Chinese medical practices with modern digital innovations.

Diandong Liu is currently pursuing his Ph.D. at the School of Electronic Information and Artificial Intelligence of the Shaanxi University of Science and Technology. He has published more than ten papers and patents. His research focuses on exploring the practical applications of large language models in various fields.

Yuxin Xie received her Bachelor's degree in nursing from Nanjing University of Chinese Medicine in 2019. Currently, she is a Nurse Practitioner in the Department of Neurosurgery at Nanjing Drum Tower Hospital. Her current research interests include the care of Moyamoya disease and epilepsy.

Ren Yi is currently pursuing a Ph.D. at the School of Software, Nanjing University. His research focuses primarily on the construction technologies of domain-specific large-language models.

Jiacun Wang is a Professor of Monmouth University, USA. His research interests include machine learning, formal methods, discrete event systems, software engineering, workflow, and distributed systems in real time. He has published four books and more than 280 papers. He is the founding Editor-in-Chief of the International Journal of Artificial Intelligence and Green Manufacturing. In addition, he serves as an Associate Editor for IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE/CAA Journal of Automatica Sinica and International Journal of Communication Systems. He is a Member-at-Large of the Board of Governors of the IEEE SMC Society and has held roles as the general chair and program chair for multiple international conferences.

Youbing Xia is a professor of Nanjing University of Chinese Medicine. With a research focus on acupuncture and medical information analysis, Prof. Xia has published nearly 80 peer-reviewed articles in journals and conferences. His work focuses on the mechanisms underlying the treatment of acupuncture for infertility and the intelligent analysis of the schools of acupuncture and moxibustion. He serves on the editorial board of several journals, including Chinese Acupuncture & Moxibustion.

Peng Zhu received his Ph.D. degree in information science from Nanjing University, Nanjing, China, in 2011. He is currently a full professor of information systems at Nanjing University of Science and Technology, Nanjing, China. His research interests include intelligent systems, high-order complex networks, blockchain technology and application, hypergraph, simplicial complex, and smart medical. His articles were published in IEEE Transactions on Engineering Management, Information Fusion, Information Processing & Management, among others.