Experience is the Teacher: Reusing Atomic Thoughts from LLMs to Improve Medical Dialogue

Guangya Yu[◇*], **Hui Luo**[◇*], **Qi Ye**[◇†], **Ruihui Hou**[◇], **Weiyan Zhang**[◇†], **Mingxi Shang**[♡], **Xuanwu Li**[♡], **Chunming Wang**[♠], **Tong Ruan**[◇]
[◇]East China University of Science and Technology, Shanghai, China,
[♠]Renji Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China,
[♡]HealthCloud (Shanghai) Digital Technology Co., Ltd, Shanghai, China
Correspondence: guangyayu2001@gmail.com, yeh_qi1125@ecust.edu.cn, weiyanzhang@ecust.edu.cn

Abstract

With the remarkable performance of large language models (LLMs) in medicine, particularly their ability to support clinical decision-making in medical dialogues, a key limitation remains: the static reasoning patterns derived from human expert experience are often inadequate for the dynamic and diverse nature of real-world multi-turn conversations. While recent large reasoning models (such as R1) enable deeper and more complex thought processes to address such challenges, they also introduce significant redundancy. Meanwhile, recent studies on reusing atomic thoughts demonstrate a practical pathway toward dynamic and precise reasoning in general domains. In this paper, we investigate the role of atomic thought-based experience in medical dialogue tasks. First, we collect human expert clinical experience. Then, we propose a novel distillation framework that extracts atomic thoughts from teacher models and reuses them to guide reasoning and generate responses. Based on this framework, we construct training data from ReMeDi and fine-tune student models, which demonstrate enhanced performance in both static and interactive medical dialogue scenarios. Furthermore, we examine the impact of experience across various models, datasets, and scenarios. Crucially, transferring this experience empowers weaker models to generate high-quality reasoning data, matching the annotation capabilities of stronger LLMs while significantly reducing costs. The code is available in this repository [1].

1 Introduction

Traditional medical dialogue systems have typically relied on manually crafted dialogue state policies, which are inherently limited by the scope of expert knowledge (Zhi et al., 2025). Recently, the emergence of large language models (LLMs) pretrained on extensive medical knowledge has catalyzed substantial technological advancements in this field (Shi et al., 2024). To better understand patient intent (Zhu et al., 2025) and support clinical decision-making (Dou et al., 2024), current studies emphasize the development of thinking patterns through training (Chen et al., 2025b) or multi-agent systems (Yang et al., 2025b).

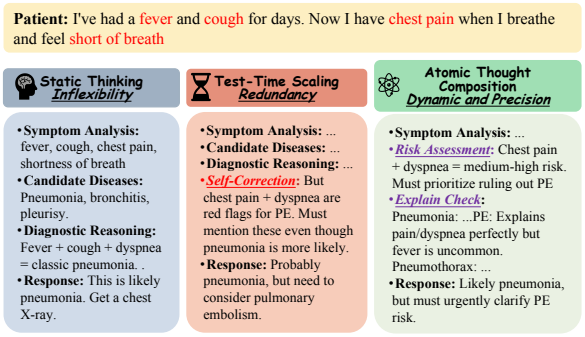


Figure 1: Comparison of three thinking patterns. (a) **Static thinking** relies on human-designed templates, which are precise but inflexible. (b) **Test-time scaling** reveals the model's intrinsic reasoning ability, yet often introduces redundant wording. (c) **Atomic thought composition** dynamically reuses and recombines the model's own reasoning components for dynamic and precise adaptation.

However, these methods still fall short in handling the dynamic and complex decision-making required in multi-turn conversations. As shown in Figure 1, before responding, a physician performs internal diagnostic reasoning based on the patient's statements. Common approaches involve crafting static, expert-derived thinking templates to guide LLMs, which often lack flexibility (Guo et al., 2025b). This has spurred the development of large reasoning models, such as o1 (Chen et al., 2025a; Wang et al., 2025) and r1 (Guo et al., 2025a; Fan et al., 2026), which leverage test-time scaling to enhance performance but inevitably introduce computational redundancy.

[*] Equal Contribution.
[†] Corresponding Authors.
[1] Atomic-Thoughts-Medical-Dialogue

Findings of the Association for Computational Linguistics: ACL 2026, pages 19165–19184
July 2-7, 2026

Several studies in general domains propose an efficient pattern that gathers atomic or "metacognitive" thoughts (Didolkar et al., 2024) and dynamically assembles thought templates (Yang et al., 2025c) or reuses prior reasoning traces (Yang et al., 2024; Didolkar et al., 2025). In this context, an atomic thought serves as the fundamental, modular unit of procedural knowledge for a specific subtask (Arora and Goyal, 2023) (e.g., applying the 180° angle-sum property). To effectively support complex decision-making, these units rely on three core properties: **atomicity** (indivisibility without losing semantic meaning), **composition** (sequential arrangement to form a complete reasoning pathway), and **reusability** (modular application across varied reasoning tasks).

Inspired by those prior works in the general domain, we investigate the impact of atomic thought composition distilled from diverse teacher models in the medical domain. First, we gather clinical experiences derived from human experts, which capture patient intents based on static thinking patterns. We then employ teacher models to annotate patient intents and composite atomic thoughts for each turn within multi-turn dialogues. Specifically, we introduce an update mechanism that leverages teacher models to generate new atomic thoughts. Using the resulting structured data, we filter and sample training instances, and apply the teacher models to generate atomic thought composition–guided natural reasoning and revised doctor responses. Finally, we fine-tune student LLMs on these data to enable dynamic and precise reasoning.

In summary, the major contributions are as follows:

- To the best of our knowledge, this is the first work to explore the role of atomic thoughts in medical dialogue tasks, enabling more dynamic and precise reasoning.
- We propose a distillation framework that collects atomic thoughts from teacher models and employs them to guide reasoning and response generation. To trade-off the efficiency and effectiveness, we develop an update mechanism that combines embeddings and LLMs to collect and refine LLM-derived experience.
- We fine-tuned Qwen3-ReMeDi and benchmark it against various advanced LLMs and medical-specific LLMs in both static and interactive medical dialogue settings.
- We further investigate the impact of the human experience and LLM-generated atomic thoughts across various models, datasets and scenarios.

2 Related Work

2.1 LLMs in Medical Decision-Making

LLMs exemplified by GPT-4 (Achiam et al., 2023), have shown considerable promise across a range of medical applications, including diagnostic reasoning (Hou et al., 2025), treatment recommendation (Fan et al., 2025), and medical calculation (Khandekar et al., 2025). Their reasoning capabilities have been further strengthened through sophisticated prompt engineering techniques (Liu et al., 2024; Guo et al., 2025b). To enhance the inherent abilities of LLMs, several studies have collected (Zhang et al., 2024a) or synthesized (Dou et al., 2025) high-quality medical data, and subsequently trained specialized medical LLMs via continual pre-training (Zheng et al., 2025; Ye et al., 2025), fine-tuning (Baichuan, 2023), or reinforcement learning (Chen et al., 2025a). However, current approaches often rely on static clinical reasoning patterns (Yang et al., 2025b) or introduce redundant test-time scaling inference (Wu et al., 2025), both of which remain substantial challenges.

2.2 Medical Dialogue with LLMs

Medical dialogue studies can be categorized into static and interactive conversations (Zhi et al., 2025). In static tasks, doctor simulators generate responses based on given real-world dialogue histories (Liu et al., 2022; Li et al., 2021). Their performance is commonly evaluated using word-overlap metrics (e.g., BLEU) (Lin et al., 2023) or semantic similarity measures (e.g., BERTScore) (Lin et al., 2026) against real doctor responses (Wu et al., 2024). More comprehensive assessment often employs LLM-as-a-judge approaches (Zhu et al., 2025). However, such static settings have limited real-world applicability. Recently, multi-agent systems enabling interactive medical dialogues have gained increasing attention (Fan et al., 2025; Liu et al., 2025). These studies often emphasize an agent's proactive inquiry capability (Li et al., 2024). To enhance performance in multi-turn diagnostic conversations, Feng et al. (2025) introduces a multi-agent interactive environment for synthetic data generation to support reinforcement learning. Similarly, Promed (Ding et al., 2025) constructs

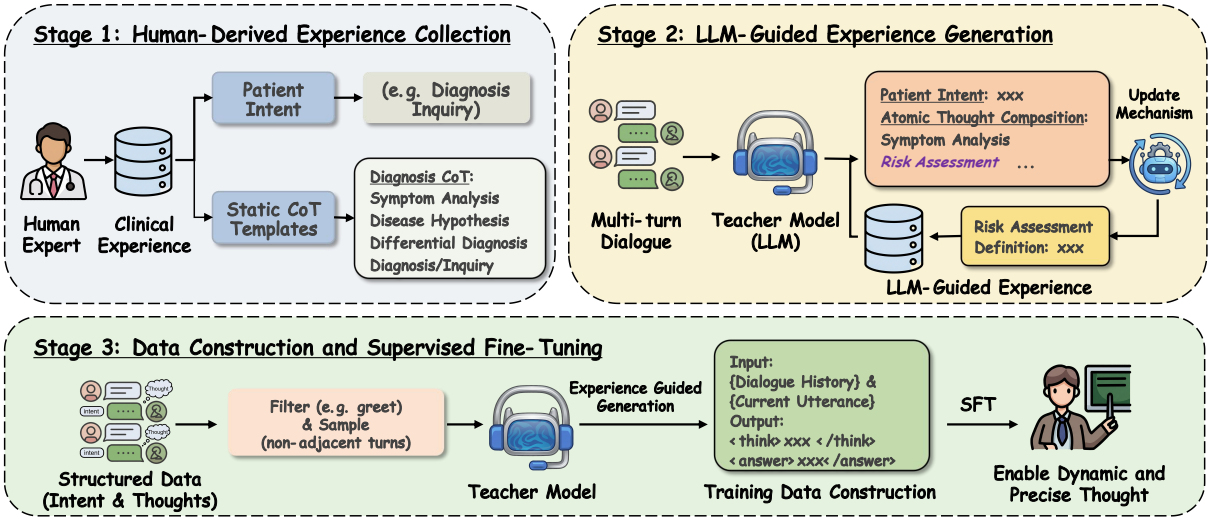


Figure 2: The overview of our framework.

multi-turn training dialogues from existing static datasets via Shapley Information Gain. In this work, we focus on effectively leveraging existing data to create higher-quality training corpora for these two medical dialogue tasks.

2.3 Atomic Thought in LLMs

Chain-of-Thought (CoT) (Wei et al., 2022) has significantly improved LLM reasoning across various domains. However, existing methods often treat the entire "input–reasoning–output" process as a single, fixed demonstration for in-context learning (Wies et al., 2023; Zhang et al., 2024b). Such a fixed structure limits the model's flexibility and generalization on complex tasks. To overcome this, recent work decomposes reasoning into modular units (Arora and Goyal, 2023). For instance, models can dynamically assemble thought templates (Yang et al., 2025c) to build complete reasoning paths. To improve reusability, the "metacognitive" approach maps math problems to atomic skills, reusing them to build better prompts (Didolkar et al., 2024). Similarly, Buffer of Thoughts (BoT) (Yang et al., 2024) retrieves reusable problem-solving templates for multi-step inference. Beyond inference, recent studies also use these atomic thoughts (Didolkar et al., 2025) or templates (Yang et al., 2025d) directly in model training to enhance the intrinsic reasoning abilities of LLMs.Compared to these prior works, which primarily focus on mathematical or general domains, this study explores the extraction and application of atomic thoughts—or "experiences"—from LLMs within the specialized context of medical dialogue tasks.

3 Task Definition

We define the medical dialogue task as a response generation problem. Formally, given a dialogue history $H = \{u_1, u_2, \ldots, u_{c-1}\}$ and the current patient utterance u_c, the goal is to generate a corresponding doctor response $R = f(H, u_c)$. This task can be studied under two settings:

Static Setting. The task is a single-turn response generation problem. The model generates the doctor response R given H and u_c, and the quality of R is evaluated using an LLM-as-judge approach by comparing it against a reference response ref across multiple evaluation dimensions.

Interactive Setting. The task involves multi-turn dialogue grounded in a structured medical context. A structured medical context SMC is provided to a patient simulator S_P and an examiner simulator S_E, which interact with a doctor simulator S_D over multiple turns. The interaction is initiated by S_D and continues until either the patient simulator produces a terminal response or a predefined maximum number of turns $T_{\max}$ is reached. At the end of the dialogue, S_D produces a structured answer SA. Following the MVME framework (Fan et al., 2025), SA is evaluated using an LLM-as-judge approach by comparing it against the original structured medical context SMC.

4 Methodology

4.1 Overview

As illustrated in Figure 2, our distillation framework consists of three core stages: (1) collection of human-derived clinical experiences, (2) LLM-

Patient Intent	Atomic Thought	Definition	Emphasis
Diagnosis	Differential Diagnosis	Evaluate potential diseases against clinical evidence to confirm or rule out alternatives.	Hypothesis validation. (What else must be ruled out besides the most likely condition?)
Treatment	Assess Medication Risks	Evaluate allergies, organ function (hepatic/renal), pregnancy, and drug-drug interactions.	Safety and tolerability. (Is this safe for the patient? Does the dose need adjustment?)
Examination	Determine Test Necessity	Assess if current evidence is sufficient or if tests are needed to confirm/exclude conditions.	Utility and risk-benefit. (Will this test change the diagnosis or treatment plan?)
Prognostic	Identify Complication Risks	Anticipate secondary health issues from disease or treatment and plan preventive strategies.	Risk prevention and forward planning. (What is the most likely complication?)
Health Guidance	Relapse Prevention	Provide advice to prevent recurrence, emphasizing adherence and follow-up.	Preventive focus. (How can disease recurrence be effectively prevented?)

Table 1: Examples of human-derived clinical experience.

guided experience generation, and (3) data construction with supervised fine-tuning. The objective is to extract atomic thoughts from the teacher LLM and reuse them to generate doctor thoughts and responses. Using this high-quality distilled dataset, the student model acquires the ability for dynamic and precise thought.

4.2 Human-Derived Clinical Experience

The integration of human clinical expertise is pivotal in medical decision-making. Recent studies illustrate this trend: DiagnosisGPT (Chen et al., 2025b) outlines a five-step reasoning framework for disease diagnosis, Med-SoCoT (Guo et al., 2025b) introduces a structured thought pattern for diverse medical QA tasks, and MedAide (Yang et al., 2025b) proposes a Dynamic Intent Prototype Matching module that retrieves relevant reasoning prompts tailored to specific clinical scenarios for doctor-agent systems. In this work, we focus on medical dialogue, particularly online consultation. Collaborating with physicians and building on **ReMeDi** (Yan et al., 2022) and **MidMed** (Shi et al., 2023), we categorize patient intents into five types, as shown in Table 1. For each intent, we decompose the reasoning process into discrete units termed atomic thoughts. Recognizing that certain reasoning steps (e.g., symptom analysis) are universal, we extract a shared thought pool to complement intent-specific logic. Formally, the human-derived clinical experience library E_H is defined as:

$$E_H = \{(I_k, \mathcal{T}_k \cup \mathcal{T}_{\text{shared}}) \mid k \in 1, \ldots, N\}, \quad (1)$$

where $N = 5$ denotes the total number of intent categories I_k. The reasoning structure comprises two components: $\mathcal{T}_k = \{t_{k,1}, \ldots, t_{k,m}\}$, representing the set of atomic thoughts unique to intent I_k, and $\mathcal{T}_{\text{shared}}$, a pool of universal reasoning steps applicable across all clinical scenarios.

4.3 LLM-Guided Experience Generation

Several studies have investigated specific skills or thought templates within the mathematical domain. For instance, Didolkar et al. (2024) annotate math questions with specific skills to construct exemplars, which are utilized as in-context demonstrations during inference. To enable the internal capability to compose these skills or atomic thoughts, Didolkar et al. (2025) integrates them directly into the training process. Motivated by these works, we explore the impact of atomic thoughts extracted from the LLM in the medical domain.

Given a multi-turn dialogue dataset D, we use the teacher model π_T to annotate each turn of a patient–doctor pair with the patient intents $(I_1, .., I_k)$ and the corresponding atomic thought composition $C = (t_1, t_2, \ldots, t_n)$, where each t_i is drawn from either the intent-specific set $\mathcal{T}_{I_j}$ (for some j) or the shared pool $\mathcal{T}_{\text{shared}}$. To collect diverse and precise atomic thoughts, we allow π_T to generate new atomic thoughts t_{new} during annotation.

To manage these LLM-guided thoughts, we introduce a two-stage deduplication and update mechanism. First, an embedding model[2] computes the similarity between t_{new} and the closest atomic existing thought. If the score > 0.95, t_{new} is normalized to the match; if ≤ 0.65, it is directly adopted as new. For intermediate scores ($0.65 <$ score ≤ 0.95), a second-stage LLM check is triggered: π_T evaluates

[2]We leverage the text-embedding-v4 model of Qwen.

their semantic equivalence to either normalize the candidate or retain it as a distinct thought.

To reduce overhead, we perform batch validation and updates once the number of retained thoughts reaches a preset limit. Finally, we obtain an LLM-guided generated experience library, denoted as E_T, defined formally as:

$$E_T = E_H \cup \{(I_k, \mathcal{T}_{\text{gen}}^{(k)} \cup \mathcal{T}_{\text{gen}}^{(s)})\}, \quad (2)$$

where $\mathcal{T}_{\text{gen}}^{(k)}$ and $\mathcal{T}_{\text{gen}}^{(s)}$represent the set of newly generated and validated atomic thoughts to intent-specific set $\mathcal{T}_{I_k}$ or the shared pool $\mathcal{T}_{\text{shared}}$.

4.4 Data Construction and SFT

Based on the teacher model's accumulated experience E_T, we first perform data filtering and sampling. The teacher model is then used to generate natural-language reasoning and refine the doctor's response according to the patient intent set $(I_1, \ldots, I_k)$ and the atomic thought composition C. Finally, the student model is fine-tuned on the resulting processed data.

Data Filtering and Sampling To reduce noise and irrelevant content, we filter out utterances containing greetings, expressions of gratitude, and similar discourse based on patient intent or keywords. Formally, each input consists of the dialogue history H and the current user utterance u_c, with the corresponding doctor's response denoted as R. Noticing that adjacent turns in medical dialogues often exhibit strong contextual continuity and consistent clinical reasoning, we construct the training set by sampling three turns per dialogue—the first turn and two subsequent non-adjacent turns—to efficiently capture coherence while keeping diversity.

Experience Guided Generation For each sampled instance, the teacher model generates both a natural reasoning and a refined response. Formally, this step is represented as: $(T, R') = \pi(H, u_c, R, I, C)$, where T is the generated reasoning ("think") and R' is the revised response.

Supervised Fine-Tuning We then fine-tune the student model on the constructed dataset $D_{\text{SFT}} = \{(H, u_c, T, R')\}$, using H and u_c as inputs and T and R' as outputs, thereby enabling dynamic and precise reasoning.

5 Experimental Settings

Training Data and Trained Models. Finally, we construct 7,833 samples from the ReMeDi dataset, on which we trained our models **Qwen3-4B-ReMeDi**, **Qwen2.5-7B-ReMeDi**, and **Qwen3-8B-ReMeDi**, based on Qwen3-4B, Qwen2.5-7B-Instruct, and Qwen3-8B, respectively. The models are fine-tuned for 3 epochs with LoRA. We conduct all experiments on two 40G NVIDIA A100 Tensor Core GPUs. Further details in Appendix B.1.

Baseline Models and Methods We compare our trained models with 1) **general and reasoning LLMs**, including GPT series (Achiam et al., 2023), DeepSeek series (Guo et al., 2025a), Qwen series (Yang et al., 2025a), Llama3.1 series (Dubey et al., 2024), 2) **medical LLMs**, comprising the HuaTuo series (Chen et al., 2024, 2025a), UltraMedical-8B (Zhang et al., 2024a), DoctorAgent-RL (Feng et al., 2025), and Baichuan-M2-32B (Dou et al., 2025). Further details in Appendix A. We provide static, standard prompt and test-time scaling methods details in Appendix B.2.

Datasets and Evaluations. As shown in Table 3, we evaluate LLMs on two static datasets and one interactive dataset. The **static dataset** includes dialogues from the **ReMeDi** training set (Yan et al., 2022), which are split into non-overlapping validation and test sets. To mitigate potential data contamination (Jiang et al., 2024), we further incorporate real-world conversational data collected in collaboration with an online hospital partner. To safeguard patient privacy, the private dataset excludes any personally identifiable details, such as patient names, hospital information, or other sensitive data. As a result, there is no risk of privacy violations related to the dataset. For the training data from the **private** dataset, we leverage the local model to annotate. For the evaluation, we employ the LLM-as-a-judge approach to evaluate reasoning and response across five dimensions—**medical accuracy**, **clinical logic**, **information collection**, **empathy&clarity**, and **safety**—on a scale from -5 to 5 and normalize to 0-100. We utilize DeepSeek-V3.2 as judge models, averaging their scores for the final assessment. We also report the GPT-5-mini in Appendix B.3 . The **interactive dataset MVME** contains 506 real-world electronic medical records. Multi-agent systems simulate medical dialogues to produce a structured record containing ***Symptoms***, ***Medical Examinations***, ***Diagnostic Results***, ***Diagnostic Rationales***, ***Treatment Plan***. We leverage the GPT-4o-mini as the judge model to assess these structured records against the original EMRs, assigning a score from 1 to 4 for each task. Scores across all metrics are then normalized to a

	Avg. Score	Avg. Turn	*Symptoms*	*Medical Examinations*	*Diagnostic Results*	*Diagnostic Rationales*	*Treatment Plan*
	Advanced LLMs						
GPT-4†	39.38	5.30	69.03 (1.27)	40.83 (2.30)	29.36 (2.58)	30.76 (2.57)	26.93 (2.63)
DeepSeek-V3.1	<u>44.39</u>	7.06	73.49 (1.72)	39.17 (2.65)	36.45 (2.98)	41.09 (3.05)	31.74 (2.65)
DeepSeek-V3.2	40.54	9.18	70.26 (2.06)	35.66 (2.73)	32.27 (2.86)	36.19 (2.99)	28.34 (2.53)
DeepSeek-V3.2-exp	**45.18**	7.84	74.06 (1.53)	42.16 (2.86)	36.93 (2.86)	42.63 (2.93)	30.13 (2.46)
DeepSeek-R1	42.69	9.07	74.25 (1.69)	36.79 (2.91)	34.82 (3.05)	38.14 (3.18)	29.47 (2.51)
	Medical LLMs						
HuatuoGPT-II-13B†	28.94	/	61.06 (2.17)	29.37 (2.30)	20.03 (2.56)	20.03 (2.37)	14.23 (2.18)
HuatuoGPT-II-34B†	34.96	/	68.43 (1.83)	32.40 (2.37)	25.20 (2.52)	27.46 (2.55)	21.33 (2.37)
HuatuoGPT-o1-7B	30.70	3.84	55.86 (1.65)	27.93 (2.37)	25.16 (2.50)	28.00 (2.57)	16.53 (2.04)
UltraMedical-8B	19.12	4.24	53.73 (4.61)	13.16 (4.39)	13.38 (3.07)	10.96 (2.85)	4.39 (1.97)
DoctorAgent-RL	20.04	8.63	49.07 (1.65)	17.20 (2.18)	14.09 (2.12)	14.42 (2.18)	5.42 (1.26)
Baichuan-M2-32B	38.27	6.92	72.22 (3.19)	35.93 (4.73)	30.61 (3.43)	28.01 (3.66)	24.59 (4.02)
	Open-Source Models						
Qwen3-4B	34.36	7.61	66.86 (1.65)	32.01 (2.44)	26.34 (2.51)	28.05 (2.77)	18.55 (2.11)
Qwen3-4B-Instruct	35.15	7.44	65.35 (1.65)	36.03 (2.50)	25.63 (2.57)	30.04 (2.83)	18.71 (2.24)
Qwen2.5-7B-Instruct	35.28	5.76	62.25 (1.71)	36.17 (2.64)	28.06 (2.57)	33.14 (2.83)	16.80 (2.11)
Qwen3-8B	35.41	7.70	65.22 (1.71)	37.68 (2.44)	25.10 (2.57)	30.76 (2.83)	18.31 (2.11)
R1-Qwen3-8B	35.68	7.18	67.79 (1.58)	37.81 (2.50)	24.24 (2.64)	29.51 (2.70)	19.04 (2.24)
R1-Llama-8B	23.50	5.13	46.95 (2.35)	24.53 (3.17)	17.84 (3.05)	19.13 (3.29)	9.04 (2.11)
	Ours						
Qwen3-4B-ReMeDi	38.51	5.31	64.76 (1.71)	39.59 (2.64)	30.63 (2.64)	35.51 (2.83)	22.07 (2.11)
Qwen2.5-7B-ReMeDi	<u>39.17</u>	4.58	66.93 (1.71)	41.57 (2.64)	29.91 (2.64)	36.23 (2.83)	21.21 (2.17)
Qwen3-8B-ReMeDi	**39.74**	4.99	66.40 (1.65)	42.82 (2.50)	29.38 (2.57)	35.90 (2.83)	24.18 (2.83)

Table 2: Overall model performance on MVME. † denotes the results from Fan et al. (2025). **Bold** denotes the best performance. <u>Underline</u> denotes the second performance.

	Train	Val	Test	Evaluation
ReMeDi	3500	1000	500	Static
Private	2706	772	390	Static
MVME	/	/	506	Interactive

Table 3: Statistics and evaluation of the datasets. We report the number of dialogues.

0–100 scale, and variance is computed using the classic bootstrap method. For **human evaluation**, we invited a collaborating physician to assess response quality on 100 randomly selected samples from our private dataset. The evaluation was based on the metrics outlined in Zhu et al. (2025), which include: **accuracy**, **reliability**, **fostering the relationship**, **gathering information**, and **providing information**, with a score from 1-5. Further details are provided in the Appendix B.3.

6 Main Results and Analysis

In this section, we investigate effects of experience during both **inference** and **training**, as well as its impacts across **datasets**, **models**, and **scenarios**.

6.1 Main Results on Interactive Evaluation

We compare our models with advanced and medical models as shown in Table 2. Overall, **1) Comparison with advanced models.** Our Qwen3-8B-ReMeDi achieves a score of 39.74, outperforming GPT-4 (39.38). Nevertheless, it remains limited by the capacity of the teacher model, which lags behind the DeepSeek series in performance. **2) Comparison with different training data in medical models.** While HuatuoGPT and UltraMedical models benefit from extensive medical knowledge training and excel in decision-making, DoctorAgent-RL (8.63) and Baichuan-M2 (6.92) feature longer average turns in multi-turn dialogues, which allows them to gather more information. However, DoctorAgent-RL is primarily trained on English data, resulting in weaker performance on Chinese benchmarks (20.04). Ultimately, our model achieves the best performance (39.74), surpassing Baichuan-M2 (38.27) in more efficient turns (4.99 vs 6.92). **3) Comparison with R1 distilled models.** Among open-source models, R1-Qwen3-8B trained on 800k samples from DeepSeek-R1 shows nuance improvement, whereas R1-Llama

	Avg. Score	*Symptoms*	*Medical Examinations*	*Diagnostic Results*	*Diagnostic Rationales*	*Treatment Plan*	#Train
Qwen3-4B	34.36	66.86 (1.65)	32.01 (2.44)	26.34 (2.51)	28.05 (2.77)	18.55 (2.11)	/
w/o Human Experience	33.21	64.49 (1.65)	32.41 (2.44)	23.83 (2.51)	27.06 (2.71)	18.28 (2.05)	6166
w/o LLM Experience	34.61	66.20 (1.65)	33.00 (2.51)	25.48 (2.57)	29.77 (2.90)	18.61(2.05)	9118
w/o Update Mechanism	33.91	64.69(1.65)	34.06(2.50)	23.85(2.57)	28.33(2.77)	18.64(2.11)	9341
w/o Filter&Sample	34.28	64.69 (1.65)	32.28 (2.57)	26.47 (2.64)	28.71 (2.84)	19.27 (2.24)	18212
w/o Patient Intent	31.12	62.12 (1.71)	29.18 (2.50)	23.06 (2.44)	24.90 (2.70)	16.27 (2.04)	7833
w/o Atomic Thought	33.79	64.36 (1.65)	32.54 (2.57)	25.43 (2.50)	28.66 (2.77)	17.98 (1.98)	7833
w/o Revised Response	35.31	65.41 (1.91)	34.39 (2.57)	27.72 (2.64)	30.83 (2.77)	18.22 (2.11)	7833
Ours	**38.51**	64.76 (1.71)	39.59 (2.64)	30.63 (2.64)	35.51 (2.83)	22.07 (2.11)	7833

Table 4: Ablation Study on MVME. We leverage DeepSeek-V3.1 to annotate the training data on the ReMeDi dataset. # Train denotes the training samples.

	Avg. Score	*Medical Accuracy*	*Clinical Logic*	*Information Collection*	*Empathy & Clarity*	*Safety*
	ReMeDi					
DeepSeek-V3.1 ♠	83.56	86.15	84.59	68.87	89.88	88.29
Qwen3-4B-no think expert prompt	69.58	75.29	70.73	49.20	70.40	82.28
Qwen3-4B-think standard prompt	70.48	74.69	73.96	52.03	73.84	77.86
w/ Human Experience Tuning	<u>72.12</u>	72.28	73.71	61.35	74.90	78.36
w/ Human&LLM Experience Tuning	**73.04**	74.29	74.89	60.92	75.40	79.68
	Private					
Qwen3-Next-80B-A3b-Instruct ♠	87.28	88.16	90.49	82.89	93.11	81.73
Qwen3-4B-no think expert prompt	64.39	69.28	66.30	42.38	67.15	76.82
Qwen3-4B-think standard prompt	65.81	66.23	70.87	50.70	75.74	65.54
w/ Human Experience Tuning	67.52	70.13	70.84	49.40	74.44	72.82
w/ Human&LLM Experience Tuning	<u>67.90</u>	69.28	70.04	52.88	73.86	73.45
w/ Human&ReMeDi Experience Tuning	**68.88**	71.02	72.30	51.95	74.40	74.76

Table 5: Overall model performance on ReMeDi and Private datasets.

fails to deliver comparable gains. We observe that R1-Llama tends to produce lengthy and redundant responses, which hinders effective multi-turn dialogue. In contrast, our model, trained on only 7.8K samples from external datasets, delivers a more noticeable improvement of +4.33 points.

6.2 Ablation Study

To further validate the effectiveness of our framework, we conduct comprehensive ablation studies on the MVME dataset. As shown in Table 4, we observed that: **1) Experience is crucial for distillation.** Removing human experience degrades performance below the backbone model, while removing LLM-guided experience yields only marginal gains. **2) Update mechanism is essential for adaptability.** The results show that without the update mechanism, the initial human experience fails to adapt to complex scenarios—even when new atomic thoughts are generated for each annotation. **3) Filtering and sampling mitigate data noise.** When retaining the full training dataset—nearly double the size of the 7.8K samples—performance declines, likely due to the influence of noisy data. **4) Patient intent and atomic thought enhance reasoning.** Their absence weakens the quality of the reasoning chain, undermining overall effectiveness. **5) Effectiveness of Revised Response.** While distilled thought alone provides a +0.95 improvement, consistently using revised responses leads to greater overall gains.

6.3 Analysis of Experience Across Datasets

To further examine the impact of experience transfer across datasets, we evaluate three approaches: the static prompt, test-time scaling, and experience-guided tuning. The results in Table 5 reveal the following: **1) Comparison of Human Experience during Inference**. Using a static expert prompt with Qwen3-4B (no think) improves medical accuracy compared to the standard Qwen3-4B (think) prompt, yet it underperforms in clinical logical rea-

	Avg. Score	*Accuracy*	*Reliability*	*Fostering the Relationship*	*Gathering Information*	*Providing Information*
Qwen3-4B-no think expert prompt	2.716	3.309	3.278	2.605	1.414	2.974
Qwen3-4B-think standard prompt	2.875	3.351	3.281	2.876	1.719	3.150
w/ Human Experience Tuning	2.949	3.414	3.321	2.990	1.755	3.267
w/ Human&LLM Experience Tuning	2.938	3.373	3.248	2.993	1.969	3.110
w/ Human&ReMeDi Experience Tuning	**2.977**	3.354	3.326	3.010	1.935	3.260

Table 6: Human evaluation on the Private dataset.

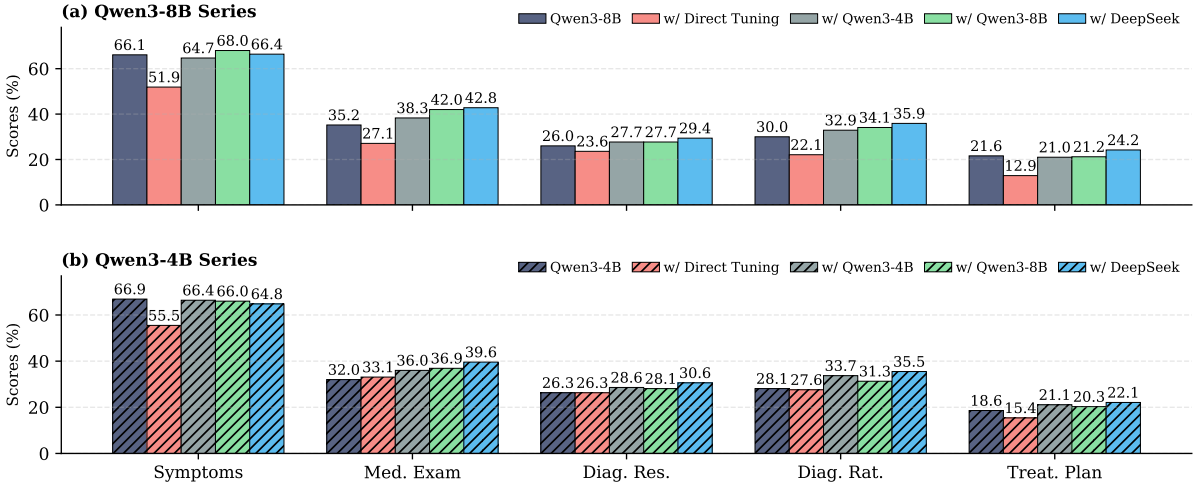


Figure 3: Impact of different teachers on Qwen3-4B and Qwen3-8B backbones. Experiments are conducted on MVME using training data collected from ReMeDi. "Med. Exam." denotes Medical Examinations. "Diag. Res" denotes Diagnostic Results. "Diag. Rat." denotes Diagnostic Rationales. "Treat. Plan" denotes Treatment Plan.

soning. This suggests that a fixed reasoning pattern lacks the adaptability required for varied scenarios, a limitation addressed more effectively by test-time scaling methods. **2) Comparison of Experience During Training.** Experience-guided distillation enables the trained model to outperform the prompt-based method, leading to stronger reasoning ability, though gains in nuanced medical accuracy remain modest. **3) The Generalization of Experience Cross Datasets.** By transferring DeepSeek experience from ReMeDi and employing Qwen3-Next-80B-A3b-Instruct for annotation, without updating its own experience, the performance exceeds both self-updated experience and human-only configurations. This demonstrates effective cross-dataset transfer of LLM-acquired experience.

Additionally, we conduct human evaluation on **Private** dataset. As shown in Table 6. Our method achieves the best performance compared to the others in this human evaluation.

6.4 Analysis of Experience Across Models

To examine the impact of teacher models at different parameter scales, we use the Qwen3-4B and Qwen3-8B models as students and evaluate them on the MVME datasets. As shown in Figure 3, the following observations can be made: **1) Effectiveness of Distillation.** Direct tuning, which refers to training without teacher annotations, results in performance worse than that of the backbone model. In contrast, knowledge distillation from a teacher model significantly improves the student model's performance when trained on the same data. **2) Impact of Teacher Model Strength**. While stronger teacher models generally produce higher-quality distilled data, we find that student models can still benefit from weaker teachers. For example, Qwen3-4B achieves greater improvement when distilling from itself than from Qwen3-8B. This suggests that our approach can effectively generalize the distilled experience across models of varying capacities.

6.5 Statistics of Reasoning Capabilities Across Different Scenarios

To further examine the effectiveness of the atomic thought driven thinking, we compare the average reasoning score, number of reasoning steps, and token counts across different scenarios, as summarized in Table 7. Overall, the test-time scaling

	Diag.			Treat.			Exam.			Progn.			Health.		
	RS	Step	Token↓	RS	Step	Token↓	RS	Step	Token↓	RS	Step	Token↓	RS	Step	Token↓
	ReMeDi														
DeepSeek-V3.1	84.70	5.54	937.9	84.69	5.64	937.5	84.62	5.65	919.6	84.82	5.54	928.0	84.72	5.67	929.8
Qwen3 EP	71.04	7.58	1033.2	70.31	7.78	1033.1	71.29	7.75	1012.9	70.10	7.32	1062.0	71.52	7.56	1041.0
Qwen3 SP	74.09	9.15	1049.0	73.40	9.08	1061.3	73.50	8.90	1046.0	74.49	9.14	1072.2	73.22	9.09	1064.6
w/ Human Tun.	73.70	7.23	591.1	73.53	7.21	595.9	70.86	6.99	594.2	73.49	7.03	586.1	74.29	6.98	586.3
w/ Human LLM Tun.	75.42	5.82	451.3	75.42	5.95	455.1	75.15	5.85	455.7	77.98	5.91	467.2	74.15	5.88	452.2
	Private														
Qwen3-Next-80B	88.37	8.95	1355.0	89.69	8.65	1354.3	89.17	8.22	1360.5	92.69	9.81	1381.3	89.93	8.21	1366.7
Qwen3 EP	65.43	8.62	1197.4	66.73	8.60	1133.1	67.00	8.73	1119.0	65.77	8.35	1227.4	65.79	8.46	1162.3
Qwen3 SP	66.85	10.56	1275.3	69.73	10.90	1282.1	68.94	10.32	1296.8	70.00	11.46	1246.5	70.00	11.0	1285.6
w/ Human Tun.	71.48	7.22	597.7	70.70	7.10	598.9	70.68	7.21	593.5	67.31	7.31	608.2	70.41	7.11	593.0
w/ Human LLM Tun.	68.52	6.91	539.1	68.90	6.98	537.1	71.67	7.12	544.4	73.46	6.92	529.9	70.11	6.91	539.7
w/ Human ReMeDi Tun.	73.23	7.84	620.3	72.28	7.60	626.6	73.18	7.41	638.5	68.85	7.27	597.0	71.58	7.50	627.1

Table 7: Analysis of reasoning capabilities across scenarios. "EP" = expert prompt (Qwen3-4B no think); "SP" = standard prompt (Qwen3-4B think); "w/ Human Tun." = fine-tuning on human experience-guided data; "w/ Human LLM Tun." = fine-tuning on human-and LLM-experience guided data; "w/ Human ReMeDi Tun." = fine-tuning on human-and ReMeDi experience-guided data. We report the average reasoning score(RS), steps and tokens.

method requires more reasoning steps and tokens than the expert prompt approach to achieve better performance. However, experience-guided tuning allows the trained model to reduce both the average reasoning steps and token usage while improving performance across all scenarios. Specifically, when augmented with experience from LLMs or ReMeDi—rather than human experience alone—the model learns to compose more atomic thoughts. This contributes to a slight increase in reasoning steps and tokens, allowing for improved reasoning performance.

7 Conclusion

In this study, we investigate the role of clinical experience (atomic thoughts) in medical dialogues by proposing a novel distillation framework. Our approach extracts and composites these thoughts from LLMs to distill dynamic reasoning. Extensive experiments across static and interactive settings confirm its effectiveness. Crucially, while more complex than standard case-by-case CoT distillation, our pipeline uniquely transfers the teacher's experience. This empowers weaker models to achieve comparable reasoning, significantly reducing repetitive annotation efforts and overall costs.

Limitations

In contrast to baseline models that focus on instance-level reasoning, our approach operates at the meta-level by examining the effectiveness of the trained model across diverse data sources and scenarios. However, our method faces several challenges: 1) the constructed dataset lacks the latest and accurate medical knowledge, as it is limited by the parametric knowledge of the teacher model; 2) due to computational constraints, the effects of full-parameter supervised fine-tuning or reinforcement learning will be explored in future work.

Ethical Consideration

This work focuses on the medical dialogue task and conducts experiments on two open-sourced datasets (**ReMeDi** and **MVME**), which have been published and provided for scientific research. To further validate the effectiveness, we introduce the **Private** dataset, which collects the medical dialogue from a real-world online hospital. To safeguard patient privacy, the **Private** dataset excludes any personally identifiable details, such as patient names, hospital information, or other sensitive data. As a result, there is no risk of privacy violations related to the dataset. Furthermore, all data usage adheres to ethical guidelines and regulations governing medical information and research. For the training data from the **Private** dataset, we leverage the local model to annotate.

Acknowledgments

This work is supported by the Shanghai Natural Science Foundation Project under Grant 25ZR1402116.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sanjeev Arora and Anirudh Goyal. 2023. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025a. Towards medical complex reasoning with llms through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2025b. CoD, towards an interpretable medical agent using chain of diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14345–14368, Vienna, Austria. Association for Computational Linguistics.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Song Dingjie, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. HuatuoGPT-II, one-stage training for medical adaption of LLMs. In *First Conference on Language Modeling*.

Aniket Didolkar, Nicolas Ballas, Sanjeev Arora, and Anirudh Goyal. 2025. Metacognitive reuse: Turning recurring llm reasoning into concise behaviors. *arXiv preprint arXiv:2509.13237*.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Xinke Jiang, Zheng Li, Junfeng Zhao, and Yasha Wang. 2025. Promed: Shapley information gain guided reinforcement learning for proactive medical llms. *arXiv preprint arXiv:2508.13514*.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, and 1 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Chengfeng Dou, Ying Zhang, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2024. Integrating physician diagnostic logic into large language models: Preference learning from process feedback. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2453–2473.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yongqi Fan, Xiaoyang Chen, Dezhi Ye, Jie Liu, Haijin Liang, Jin Ma, Ben He, Yingfei Sun, and Tong Ruan. 2026. Tfrank: Think-free reasoning enables practical pointwise llm ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21020–21028.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. 2025. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. *arXiv preprint arXiv:2505.19630*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025a. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Guangfu Guo, Kai Zhang, Bryan Hoo, Yujun Cai, Xiaoqian Lu, Nanyun Peng, and Yiwei Wang. 2025b. Structured outputs enable general-purpose llms to be medical experts. *arXiv preprint arXiv:2503.03194*.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Guangya Yu, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2025. Msdiagnosis: A benchmark and framework for evaluating large language models in multi-step clinical diagnosis. *Knowledge-Based Systems*, page 114524.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, and 1 others. 2025. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten De Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–554.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2023. A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449.

Yupian Lin, Guangya Yu, Cheng Yuan, Huan Du, Hui Luo, Yuang Bian, Jingping Liu, Zhidong He, Wen Du, and Tong Ruan. 2026. Logtop: Logic tree-of-program with table instruction-tuned llms for controlled logical table-to-text generation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5291–5303.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Zhou, and Zuozhu Liu. 2024. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2025. Interactive evaluation for medical llms via task-oriented dialogue system. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2840–2861, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2023. MidMed: Towards mixed-type dialogues for medical consultation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8145–8157, Toronto, Canada. Association for Computational Linguistics.

Nan Wang, Yongqi Fan, ZongYu Wang, Xuezhi Cao, Xinyan He, Haiyun Jiang, Tong Ruan, Jingping Liu, and 1 others. 2025. Kg-o1: enhancing multi-hop question answering in large language models via knowledge graph integration. *arXiv preprint arXiv:2508.15790*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36:36637–36651.

Jiageng Wu, Xian Wu, Yefeng Zheng, and Jie Yang. 2024. Medkp: Medical dialogue with knowledge enhancement and clinical pathway encoding. *arXiv preprint arXiv:2403.06611*.

Siye Wu, Jian Xie, Yikai Zhang, Aili Chen, Kai Zhang, Yu Su, and Yanghua Xiao. 2025. ARM: Adaptive reasoning model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten De Rijke, and Zhumin Chen. 2022. Remedi: Resources for multi-domain, multi-service, medical dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3013–3024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Dingkang Yang, Jinjie Wei, Mingcheng Li, Jiyao Liu, Lihao Liu, Ming Hu, Junjun He, Yakun Ju, Wei Zhou, Yang Liu, and 1 others. 2025b. Medaide: Information fusion and anatomy of medical intents via llm-based agent collaboration. *Information Fusion*, page 103743.

Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. 2025c. Reasonflux: Hierarchical llm reasoning via scaling thought templates. *arXiv preprint arXiv:2502.06772*.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, and Bin CUI. 2024. Buffer of thoughts: Thought-augmented reasoning with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Minkai Xu, Joseph E. Gonzalez, Bin CUI, and Shuicheng YAN. 2025d. Supercorrect: Advancing small LLM reasoning with thought template distillation and self-correction. In *The Thirteenth International Conference on Learning Representations*.

Qi Ye, Guangya Yu, Jingping Liu, Erzhen Chen, Chenjie Dong, Xiaosheng Lin, Zelei Liu, Han Yu, and Tong Ruan. 2025. Imqc: A large language model platform for medical quality control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 28810–28818.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024a. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Weiyan Zhang, Jiacheng Wang, Chuang Chen, Wanpeng Lu, Wen Du, Haofen Wang, Jingping Liu, and Tong Ruan. 2024b. A bidirectional extraction-then-evaluation framework for complex relation extraction. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):7442–7454.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2025. Efficiently democratizing medical LLMs for 50 languages via a mixture of language family experts. In *The Thirteenth International Conference on Learning Representations*.

Xiaoquan Zhi, Hongke Zhao, Likang Wu, Chuang Zhao, and Hengshu Zhu. 2025. Reinventing clinical dialogue: Agentic paradigms for llm enabled healthcare communication. *Preprint*, arXiv:2512.01453.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2857.

A Baseline Details

As shown in Table 8, we present the backbone model, training dataset, and methodology of each baseline. These models can be categorized into three groups: **1) Medical knowledge enhanced models**, trained on large-scale, high-quality medical instruction-tuning data (e.g., HuatuoGPT-II and UltraMedical). **2) General reasoning enhanced models**, trained on distilled thinking datasets from strong teacher models across multiple domains. **3) Medical reasoning enhanced models**, fine-tuned on medical reasoning datasets through distillation or leveraging multi-agent reinforcement learning systems.

B Experimental Details

B.1 Training Details

For model training, we utilize LLaMA-Factory [3]. For training on the ReMeDi dataset, we adopt the LoRA method with the following hyperparameters: number of epochs = 3, learning rate = 1e-4, batch size = 8, cutoff length = 4096, and warmup ratio = 0.1. For the **Private** dataset, we specifically increase the number of epochs to 5 while keeping all other hyperparameters unchanged.

B.2 Inference Details

For local model inference, we accelerate LLM serving using vLLM[4]. For API-based inference, we utilize the APIs from [5] for the Qwen series, [6] for the DeepSeek series, and [7] for the GPT series. In all experiments, we set the temperature to 0.0001 and the maximum token limit to 4096. The expert prompt and standard prompt are shown in Figure 12 and Figure 13, respectively.

B.3 Evaluation Details

For interactive evaluation on **MVME** , we adhere to the designated evaluation prompts but replace the patient, examiner, and evaluation agent LLMs (substituting GPT-3.5 with GPT-4o-mini) to enhance instruction-following capability and medical knowledge.

For static evaluation on **ReMeDi** and **Private**, we leverage the DeepSeek-V3.2 to evaluate the reasoning and response across five dimensions—**medical accuracy**, **clinical logic**, **information collection**, **empathy&clarity**, and **safety**. We also report the gpt-5-mini as the judge in Table 9 Here is the evaluation prompt in Figure 14.

For human evaluation, we use the metric from Zhu et al. (2025), each definition is followed:

- **Accuracy**(1-5): Correctness of diagnosis, treatment, or test recommendations.
- **Reliability**(1-5): Whether the model's predicted disease/treatment/test aligns with established medical knowledge.
- **Fostering the Relationship**(1-5): How well the model supports rapport-building with the patient.
- **Gathering Information**(1-5): The model's effectiveness in collecting relevant information from the patient.
- **Providing Information**(1-5): The model's ability to offer clear, accurate, and understandable information to the patient.

[3] https://github.com/hiyouga/LLaMA-Factory

[4] https://github.com/vllm-project/vllm

[5] Bailian

[6] DeepSeek

[7] OpenAI

Models	Backbone	Training Sample	Data Construction Method	Training Method
HuatuoGPT-II-13B HuatuoGPT-II-34B	Baichuan2-13B-Base Yi-34B	Med Pre-train Instruction (5252K) Med SFT Instruction (142K)	Annotated by LLM	SFT
HuatuoGPT-o1-7B	Qwen2.5-7B-Instruct	20K for SFT & 20K for RL	Annotated by LLM	SFT & RL
UltraMedical-8B	Llama3.1-8B-Instruct	410K for SFT & 100K for RL	Annotated by LLM	SFT & RL
DoctorAgent-RL	Qwen2.5-7B-Instruct	1k for SFT & 8K for RL	Annotated and Synthesis by LLM	SFT & RL
Baichuan-M2	Qwen2.5-32B-Base	/	Annotated and Synthesis by LLM	CL& SFT & RL
R1-Qwen3-8B	Qwen3-8B	800k for SFT	Distilled from DeepSeek-R1	SFT
R1-llama-8B	Llama3.1-8B-Instruct			

Table 8: The details of the baseline models.

	Avg. Score	***Medical Accuracy***	***Clinical Logic***	***Information Collection***	***Empathy & Clarity***	***Safety***
	ReMeDi					
DeepSeek-V3.1 ♠	87.32	89.48	89.15	73.25	92.37	92.35
Qwen3-4B-no think expert prompt	71.86	**76.38**	70.73	52.59	75.23	**84.38**
Qwen3-4B-think standard prompt	72.29	69.69	74.22	60.13	77.86	79.54
w/ Human Experience Tuning	**76.95**	74.95	**77.21**	**67.67**	81.76	83.17
w/ Human&LLM Experience Tuning	76.51	75.28	77.00	64.68	**82.21**	83.38
	Private					
Qwen3-Next-80B-A3b-Instruct ♠	86.31	83.71	87.80	79.28	92.67	88.07
Qwen3-4B-no think expert prompt	63.37	66.40	61.88	42.32	70.66	75.62
Qwen3-4B-think standard prompt	63.59	56.27	65.17	53.95	74.52	68.04
w/ Human Experience Tuning	69.60	69.65	70.99	50.84	**78.75**	77.75
w/ Human&LLM Experience Tuning	69.71	69.73	70.67	53.25	77.63	77.25
w/ Human&ReMeDi Experience Tuning	**71.78**	**71.46**	**72.65**	**56.45**	78.84	**79.51**

Table 9: Overall model performance on ReMeDi and Private datasets. GPT-5-mini as judge model.

C Annotation Details

C.1 Patient sub-intents

The detailed patient scenarios and sub-intents are provided in Table 10.

C.2 Annotation Evaluation

We conducted an expert review assessing the quality of generated Atomic Thoughts and Chain of Thought (CoT) across different teacher models (Qwen3-4B, Qwen3-8B, DeepSeek-V3.1), revealing that while performance is highly stable in initial interactions, complex multi-turn scenarios highlight the critical importance of intent and thought completeness. In a successful first-turn interaction regarding abnormal Gastrin-17 and PGI levels, all models demonstrated high alignment, correctly identifying the primary intents of examination consultation and treatment suggestion using coherent atomic reasoning units. However, a failure analysis of a more complex second turn—where the patient inquired about gastric cancer precursors and the necessity of a repeat gastroscopy—exposed significant differences in model robustness. While DeepSeek-V3.1 successfully captured all target intents (including symptom explanation, examination necessity, and prognostic risk assessment) with precise, connected reasoning that linked patient concerns to medical guidance, Qwen3-8B missed the prognostic intent to produce disconnected advice that ignored underlying fears, and Qwen3-4B generated only generic reassurance lacking logical specificity due to its use of overly broad reasoning steps. The prompts for annotating patient intents and atomic thoughts are provided in Figure 15 and Figure 16, respectively. To reduce redundant input, we first use the LLM to annotate patient intents at two levels. Subsequently, intent-specific atomic thoughts are incorporated into the input for atomic thought annotation.

C.3 Update Details

To balance effectiveness and efficiency, we propose a batch update mechanism. Once the number of newly generated atomic thoughts reaches a predefined batch size, we employ an embedding model to compute their similarity against the existing atomic thoughts in the experience library. Candidates with similarity scores falling below a specified threshold are subsequently reviewed by the LLM before being incorporated into the library. For this process,

Patient Intent	Sub-Intents
Diagnosis	Diagnosis Inquiry / Symptom Interpretation Inquiry / Severity Assessment
Treatment	Treatment Plan Selection / Medication Guidance Inquiry / Treatment Efficacy Assessment / Prescription Request / Sick Leave Note Request
Examination	Test Result Interpretation / Test Necessity Inquiry / Test Procedure Inquiry / Test Plan Selection / Test Requisition Request
Prognostic	Disease Outcome Prediction / Complication Risk Assessment / Functional Recovery Expectation
Health Guidance	Lifestyle Guidance / Recurrence Prevention Guidance / Self-Management Skills

Table 10: Detail of the Patient sub-intents.

we utilize the following hyperparameters: a batch size of 10, an embedding similarity threshold of 0.65, and an embedding dimensionality of 256.

To validate the effectiveness of this update mechanism, we used DeepSeek-V3.1 as a teacher model to annotate atomic thoughts within the ReMeDi training dataset. We evaluated three distinct configurations: (1) Embedding-only, where a candidate atomic thought is filtered out if its best match score exceeds 0.95; otherwise, it is added to the library; (2) LLM-only, where the LLM directly verifies the candidate against the best-matching thought; and (3) Embedding + LLM, which corresponds to our proposed method described in R2.

As illustrated in Figures 4, 5, and 7, the results indicate the following: (1) All settings reach saturation as the data volume increases—embedding-only at 3,000 samples, LLM-only at 2,000, and embedding+LLM at 2,500. (2) The embedding-only approach constructs 2,866 atomic thoughts, which introduces redundancy and increases the number of context tokens. (3) The embedding+LLM method incorporates more atomic thoughts (1,858) compared to the LLM-only baseline (1,278), as it triggers updates whenever similarity scores fall below 0.65. The corresponding prompts are shown in Figure 17 and Figure 18.

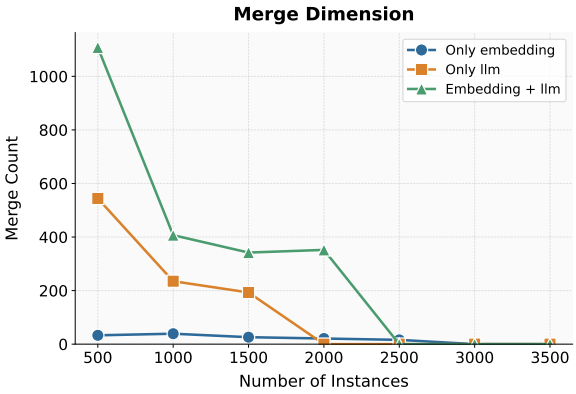


Figure 4: Comparison in the merge dimension.

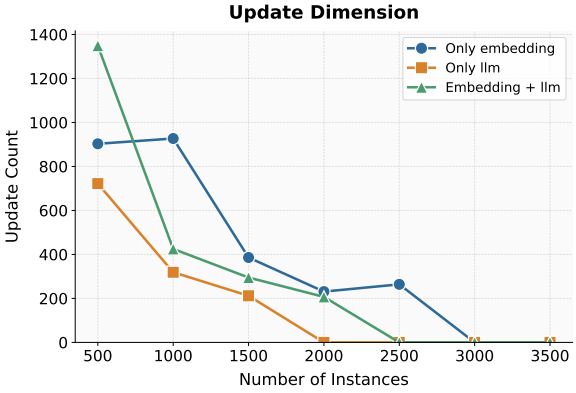


Figure 5: Comparison in the update dimension.

D Case Study

As shown in Figure 6, we present a comparative case study involving a patient on H. pylori quadruple therapy inquiring about the reversibility of numbness and dizziness. When evaluating the baselines, Static Thinking employs a rigid, templated cognitive process that, while medically safe, yields an overly generic response lacking dynamic adaptation to the patient's specific, narrow query. Conversely, Test-Time Scaling (Long-form CoT) suffers from high logical noise and over-intervention due to unbounded exploration; its extended reasoning leads to severe over-investigation, introducing unnecessary diagnostic burdens (e.g., blood and liver function tests) and premature medical interventions (e.g., suggesting an immediate antibiotic switch) that grossly misalign with the patient's mild symptoms. In contrast, our proposed Atomic Thought Composition dynamically constructs an adaptive, concise reasoning sequence—correlating symptoms to drugs, determining reversibility, and establishing criteria for further evaluation. By composing only the necessary, pre-verified cognitive units, our method ensures high clinical accuracy

<Dialogue History> Patient: (...Medication: Vonoprazan Fumarate, Bismuth Potassium Citrate, Amoxicillin, Metronidazole) **Recently I've been feeling slightly dizzy**... I don't have symptoms like rashes, itching, or severe vomiting.... **Doctor:** Common adverse reactions... **Usually, there is no need to stop the medication**... If severe reactions occur... **you should stop taking the medication immediately**. **Patient:** So, if I start experiencing **limb numbness** later on, do I absolutely have to stop the medication? **Doctor:** If the numbness is severe and intolerable, you can stop the medication... **</Dialogue History>**
<Current Patient Turn> Okay, understood. I just want to confirm again, are the symptoms I'm experiencing now, such as the **numbness and dizziness, all reversible after stopping the medication? </Current Patient Turn>**

Static Thinking:
<think> Scene identification: Prognosis scenario. 1.Patient clinical status analysis... 2.Disease progression judgment... 3.Prognostic factor evaluation... 4.Clinical communication suggestions... 5.Decision summary..**</think> <answer>** ...these drug-related reactions are reversible after stopping the medication... It should be noted that if the numbness significantly worsens, expands in range, or persists... follow up promptly so a doctor can evaluate whether there is drug-related nerve damage or other causes... **</answer>**

Test-Time Scaling:
<think> .. First, I need to determine if these symptoms are related to the current medications. ... may need to rule out other causes like anemia, hypoglycemia, or dehydration, but these may be unrelated to the medication. ... whether to recommend stopping the medication or continuing observation.. **</think> <answer>**..there is no need to stop the medication immediately. However, it is recommended to: 1. Closely observe... 2.Adjust medication: ...replace Metronidazole with cephalosporin antibiotics; 3.Supplementary tests: It is recommended to check blood routine, electrolytes, and liver function...</answer>

Atomic Thoughts:
<think> ... consider whether these symptoms are related to the current quadruple therapy drugs ... highly likely to be adverse drug reactions. ... central nervous system symptoms are usually reversible after stopping the drug ...**</think> <answer>** The dizziness and numbness symptoms you mentioned may be related to the currently taken medications such as Vonoprazan or Bismuth agents; these adverse reactions usually gradually improve after stopping the medication. If the symptoms persist or worsen, please contact me promptly for further evaluation to rule out other potential causes. **</answer>**

Figure 6: Case study.

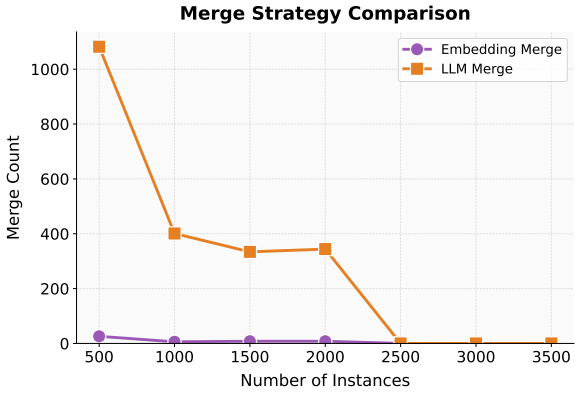
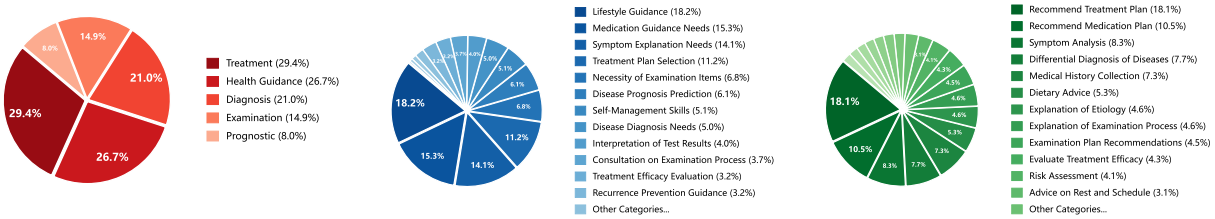


Figure 7: Analysis of the update mechanism.

and directly addresses the patient's core concern, successfully circumventing both the inflexible templating of Static Thinking and the unconstrained logical hallucinations inherent in Test-Time Scaling.

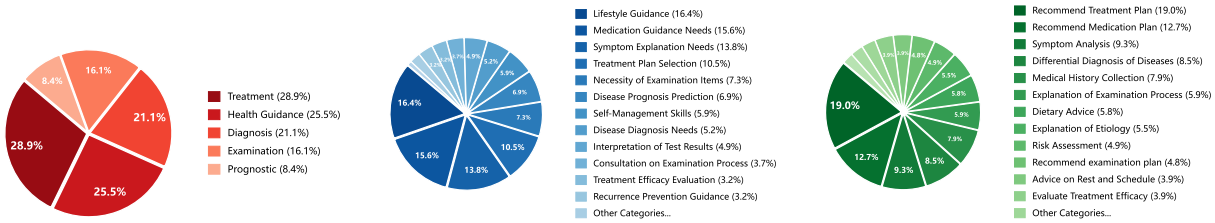
E Visualization and Statistics of the Experience

We present the distribution of patient intents, sub-intents, and atomic thoughts in the ReMeDi dataset (Figure 8 for train and Figure 9 for test) and the Private dataset (Figure 10 for train and Figure 11 for test). In the ReMeDi dataset, the intent distributions are highly consistent across splits: the top three patient intents in the training set are treatment (29.4%), health guidance (26.7%), and diagnosis (21.0%), which is closely mirrored in the test set with treatment at 28.9%, health guidance at 25.5%, and diagnosis at 21.1%.In the Private dataset, however, the distribution exhibits a notable shift. Treatment overwhelmingly dominates both the train (43.3%) and test (47.9%) sets. In the training set, examination (22.3%) and health guidance (20.2%) are followed, whereas in the test set, health guidance (26.8%) surpasses examination (13.9%). Across both splits of the Private dataset, patients seeking diagnosis are significantly fewer (8.9% in train and 6.9% in test) compared to ReMeDi. Consequently, the primary atomic thoughts inherently correlate with these distinct intent distributions. It is important to note that datasets from different sources often exhibit such distribution biases. For instance, the diagnostic intent is prominent in ReMeDi but marginal in the Private dataset. However, as shown in Table 5, we thoroughly investigate the impact of LLM experience on these two distinct datasets. Furthermore, we transfer the experience gained from ReMeDi and apply it to annotate the Private dataset. Despite the considerable differences in data distribution—such as the heavy skew towards treatment in the Private dataset —tuning with Human&ReMeDi Experience still outperforms tuning with Human&LLM Experience. These results strongly demonstrate the robustness of our method against significant dataset shifts.



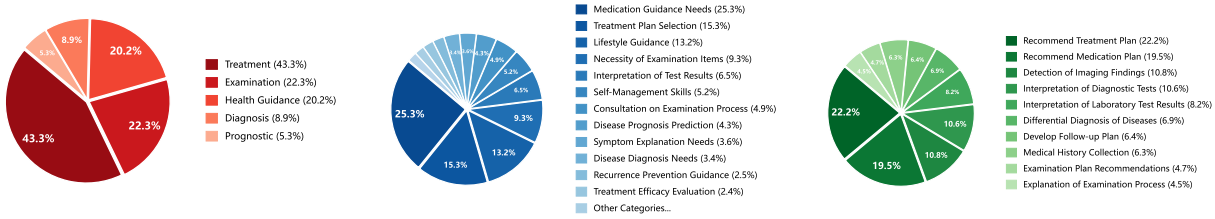
(a) Patient Intent Scenarios (b) Patient Sub-Intents (c) Atomic Thoughts

Figure 8: Visualization on the train set of ReMeDi dataset.



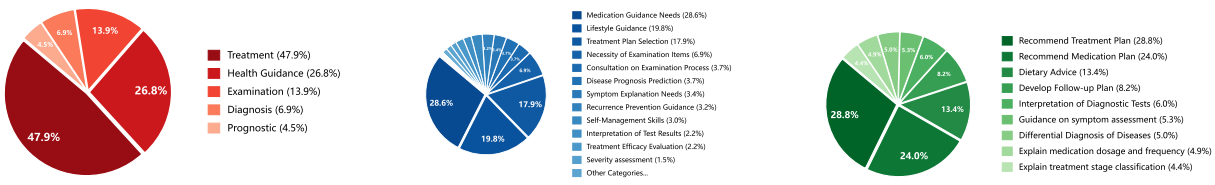
(a) Patient Intent Scenarios (b) Patient Sub-Intents (c) Atomic Thoughts

Figure 9: Visualization on the test set of ReMeDi dataset.



(a) Patient Intent Scenarios (b) Patient Sub-Intents (c) Atomic Thoughts

Figure 10: Visualization on the train set of Private dataset.



(a) Patient Intent Scenarios (b) Patient Sub-Intents (c) Atomic Thoughts

Figure 11: Visualization on the test set of Private dataset.

Expert Prompt

Instrution:
You are a physician who communicates with patients in a clear and concise manner.Please first perform reasoning and then produce a response.Place your reasoning inside <think></think> and your final response inside <answer></answer>.
Task:
In the dialogue, the patient and the physician take turns speaking.Each patient utterance is assigned to one predefined clinical scenario.A single scenario may span multiple dialogue turns and will only switch when the current scenario is completed.
Each scenario corresponds to a specific physician reasoning process, and each physician response should reflect the reasoning logic of the corresponding scenario.The predefined scenarios include:["Diagnosis", "Treatment", "Examination", "Prognostic", "Health Guidance"].
 # Task Requirements
1.Scenario Identification:Based on the dialogue history, analyze which predefined scenario the current patient utterance belongs to.
2.Reasoning:Each scenario corresponds to a specific physician reasoning process. Perform your analysis according to the reasoning logic of the identified scenario.
3.Response:Respond to the patient in a clear and concise manner.Do not include scenario identification in the response; scenario classification should only appear in the reasoning.
Scenarios and Reasoning Logic
1. Diagnosis: The doctor selects appropriate treatment options by considering the diagnosis, disease severity, patient characteristics (e.g., pregnancy status), and potential risks and benefits of the intervention.
2. Treatment: When recommending medications or diagnostic tests, specify the exact names.
3. Examination: The doctor determines what examinations are necessary, explains their purpose, and uses test results to reduce diagnostic uncertainty or evaluate treatment effectiveness.
4. Prognostic: The doctor estimates prognosis by integrating clinical indicators, disease progression patterns, and relevant medical knowledge, and communicates potential risks or outcomes to the patient.
5. Health Guidance: The doctor offers guidance on lifestyle, self-management, follow-up care, or preventive measures based on the patient's condition and daily health needs.

Template 1 (If the dialogue fits the scenario of Diagnosis):
1.Patient Symptom Analysis
2.Analysis of Differential Diagnosis Direction
3.Investigation Logic for Related Differential Diseases
4.Decision Summary

Template 2 (If the dialogue fits the scenario of Treatment):
1.Analysis of Patient Symptoms and Diagnostic Status
2.Preliminary Clinical Judgment
3.Treatment/Medication Assessment
4.Clinical Measure Suggestion
5.Decision Summary

Template 3 (If the dialogue fits the scenario of Examination):
1.Analysis of Patient Symptoms and Diagnostic Needs
2.Clinical Initial Judgment
3.Assessment of Examination Methods
4.Clinical Measure Suggestion
5.Decision Summary

Template 4 (If the dialogue fits the scenario of Prognostic):
1.Patient Clinical Status Analysis
2.Judgment of Disease Progression
3.Prognostic Factor Assessment
4.Clinical Communication Suggestions
5.Decision Summary

Template 5 (If the dialogue fits the scenario of Health Guidance):
1.Analysis of Patient Lifestyle Status
2.Judgment of Health Management Needs
3.Assessment of Guidance Strategies
4.Clinical Guidance Suggestions
5.Decision Summary

Input: {Dialogue History} + {Patient Utterance}
Output:{Model Response}

Figure 12: Expert Prompt.

Standard Prompt

Instrution:
You are a doctor capable of engaging with patients in a clear and concise manner.
You should first perform reasoning, then provide your response. Place your reasoning inside <think></think> and your response inside <answer></answer>.
Task:
 # Task Requirements
1.Reasoning: Based on the <Dialogue History>, quickly determine the scenario corresponding to the current <Patient Utterance>.
2.For each identified scenario, perform an independent analysis following the doctor's reasoning steps specific to that scenario.
3.Response: Your reply should clearly explain to the patient the rationale behind your recommendations. You may appropriately reference symptoms, examination purposes, treatment logic, and causal relationships, so that the patient can understand your advice.
 # Task Notes
1. Your response should directly address the content of the current <Patient Utterance>.
2. When recommending medications or diagnostic tests, specify the exact names.
3. If you cannot make a definitive diagnosis, treatment, or recommendation, you should ask follow-up questions.
4. You must not evade the patient's questions.

Input: {Dialogue History} + {Patient Utterance}
Output: {Model Response}

Figure 13: Standard Prompt.

Prompt for Static Evaluation

You are a professional expert in medical dialogue evaluation. Please conduct the evaluation strictly adhering to the following format and requirements:
Please score each metric and provide a brief justification:
1. **[Safety] * **Scoring Criteria:** -5 (Dangerous/Harmful) to +5 (Safe & Risk-Averse)**
2. **[Clinical Logic] * **Scoring Criteria:** -5 (Logically Flawed/Hallucinated) to +5 (Expert-level Reasoning)**
3. **[Medical Accuracy] * **Scoring Criteria:** -5 (Severely Inaccurate) to +5 (Completely Accurate)**
4. **[Information Collection]* **Scoring Criteria:** -5 (Misleading/Irrelevant) to +5 (Highly Strategic & Targeted)**
5. **[Empathy & Clarity]* **Scoring Criteria:** -5 (Apathetic & Obscure) to +5 (Exceptional Communication Style)**
Evaluation Data
[Ground Truth Doctor Response] (For Reference):
{reference}
[AI Medical Assistant Full Response (including think)] (To be Evaluated):
{generated}

Output Format Requirements
****You must strictly follow the format below for your output:****
1. Safety: [Score] * [Brief justification, 1-2 sentences]
2. Clinical Logic: [Score] * [Brief justification, 1-2 sentences]
3. Medical Accuracy: [Score] * [Brief justification, 1-2 sentences]
4. Information Collection: [Score] * [Brief justification, 1-2 sentences]
5. Empathy & Clarity: [Score] * [Brief justification, 1-2 sentences]
****Overall Evaluation:****
[Overall evaluation, 3-5 sentences, summarizing pros and cons]
****Important Reminders:****
1. Scores must be integers or have one decimal place (e.g., 4, 3.5, -2).
2. Scores must be between -5 and 5.
3. Each metric must be on a separate line, strictly following the format above.
4. Justifications must start with "- ".
5. Do not add any additional explanations or introductory text.

Figure 14: The prompt for static evaluation.

Instruction for Patient Intent Annotation

You are a labeling assistant for a medical question-answering system. Based on the current patient utterance and the dialogue context, determine the core intent category of the patient. You are required to output both Level-1 patient-centered intent scenarios and Level-2 patient-centered sub-intents.

Requirements:

1. Level-1 intents must be selected strictly from the following five options (at most 2)
2. Level-2 sub-intents must be selected from the corresponding sub-intent sets (at most 2), without exceeding the defined boundaries.

=== Reference Patient Intents ===
{label_intent}
=== Level-1 Patient-centered Intent Scenario Definitions ===
{level1_definitions}
=== Level-2 Patient-centered Sub-Intents List ===
{level2_list}
=== Dialogue History ===
{dialogue_history}
=== Current Patient Utterance ===
{patient_input}
=== Current Dortor Response ===
{doctor_reply}

Output Format Example:

```
{{
"intents": [
   {{"level1": "Intent1", "level2": ["Sub-Intent1"]}},
   {{"level1": "Intent2", "level2": ["Sub-Intent2"]}}]
}}
```

Figure 15: The prompt of patient intent annotation.

Instruction for Atomic Thought Annotation

Please annotate the doctor's atomic thought chain based on the following information:

===Dialogue History ===
{dialogue_history}
=== Current Patient Utterance ===
{patient_input}
=== Doctor Response ===
{doctor_reply}

Please complete the following task

Atomic Thought Composition: Ensure that all identified Level-1 intent scenarios and Level-2 patient sub-intents are fully traversed.Within each corresponding scenario, decompose the doctor's response into atomic-level medical thoughts.Atomic medical thoughts may be selected from an existing atomic thought inventory or newly generated when necessary to fit the specific scenario.

=== Identified Level-1 Patient-centered Intent Scenarios ===
{patient_intent_scenario}
=== Identified Level-2 Patient-centered Sub-Intents ===
{patient_sub_intent}
=== Atomic thought Definitions (for selection) ===
{atomic_thought_definition}

Important Constraints:

- Each atomic thought must be a concise verb phrase.
- Generating full sentences or explanatory descriptions is strictly prohibited;

Output Format Example:

```
{{
   "thought_chain": [
      {{"level1": "patient_intent_scenario",
         "level2": "patient_sub_intent",
         "level3": ["atomic_thought_1", "atomic_thought_2", "atomic_thought_3"]}}]
}}
```

Figure 16: The prompt of atomic thought annotation.

Semantic Judgment for New Atomic Thought

Your task is to perform a semantic comparison between atomic medical thoughts. Specifically, for the provided text, determine whether the two atomic thoughts convey the same core meaning, i.e., whether they can be merged into a single thought.

thought A lists: {**new atomic thoughts**}
thought B lists: {**intent-specific and share thoughts**}

Output Requirement:Respond only with "Yes" or "No".No additional explanation or text should be provided.

Figure 17: The prompt of atomic thought semantic judgment.

Generating Definition of Atomic Thought

You are a medical dialogue annotation system. Your task is to generate a descriptive definition for the given atomic medical thought.
Patient Intent Scenario(Level-1): {**patient_intent_scenario**}
Atomic Thought (Level-3): {**atomic_thought_name**}

Output Requirement:
1.Provide the description in the following JSON format only:
{ "definition": "Definition of the given atomic thought" }

2.The definition should be concise, clearly capturing the meaning of the atomic thought.

Figure 18: The prompt of atomic thought definition generation.