Triplet-Structured Knowledge Integration for Interactive Medical Question Answering

Zhaohan Meng[1], **Zaiqiao Meng**[1], **Siwei Liu**[2], **Iadh Ounis**[1]
[1]School of Computing Science, University of Glasgow, UK
[2]School of Natural and Computing Science, University of Aberdeen, UK

Abstract

Interactive medical question answering (QA) aims to model real-world clinical diagnosis, where a doctor gathers patient information across multiple dialogue turns. Although several approaches have been proposed for interactive medical QA, they primarily reason over dialogue history, where patient information is scattered across turns without explicit connections, hindering multi-hop reasoning. In this paper, we propose **TriMediQ**, a triplet-structured approach that converts patient responses into triplets, aggregates them into a patient-specific knowledge graph, and injects structured representations into a doctor agent via a trainable projection module. Experiments on two interactive datasets show that TriMediQ consistently improves performance over dialogue-only baselines across multiple LLM backbones, achieving up to a **10.4% absolute improvement** in the iMedQA dataset. These results demonstrate that structured knowledge representations provide an effective approach for interactive medical QA[1].

1 Introduction

Traditional medical question answering (QA) datasets are typically formulated as single-turn question–answer pairs, where existing models (Liu et al., 2020; Shang et al., 2025) have already achieved a strong performance. However, such settings fail to capture the interactive nature of real-world clinical practice, where diagnoses are formed through iterative questioning and evidence accumulation (Masic, 2022). In realistic scenarios, doctors refine diagnostic hypotheses over multiple turns by selectively eliciting missing patient information to support final decisions (Hendrycks et al., 2021). As a result, interactive medical QA has emerged as a critical task for simulating real-world diagnostic processes (Jin et al., 2024; Wang et al., 2023). One of the core challenges in interactive medical QA is the ability to explicitly connect distributed clinical facts across dialogue turns in order to perform reliable multi-hop reasoning (Li et al., 2024). For instance, a doctor-like conversational agent needs to connect a symptom mentioned early in the conversation with a medication revealed later in the dialogue, as illustrated in Fig. 1. This motivates the development of interactive QA paradigms where Large Language Models (LLMs) must actively engage in multi-turn information integration for multi-hop reasoning rather than passively responding to isolated questions.

To address the challenge, some approaches have been proposed for medical QA (Shang et al., 2025; Zhang et al., 2025; Li et al., 2024). For example, MEDIQ (Li et al., 2024) simulates a doctor–patient dialogue setting in which a doctor agent actively queries a patient simulator and reasons over accumulated dialogue logs. In addition, ProMed (Ding et al., 2025) introduces reinforcement learning with Shapley Information Gain (Winter, 2002) rewards to encourage proactive questioning. Although these approaches demonstrate the effectiveness of interactive paradigms in medical QA, a key challenge lies not only in how the doctor agent asks questions, but also in the absence of explicit structure in patient information acquired across dialogue turns. Knowledge triplets can capture clinically relevant facts as entities and relations, but when extracted from separate dialogue turns, they become fragmented and miss cross-turn dependencies. As shown in Fig. 1, facts such as symptoms and long-term medication use often appear in different turns, making it difficult for the doctor agent to perform multi-

[1]The complete code and datasets are available: https://anonymous.4open.science/r/TriMediQ-6CF4.

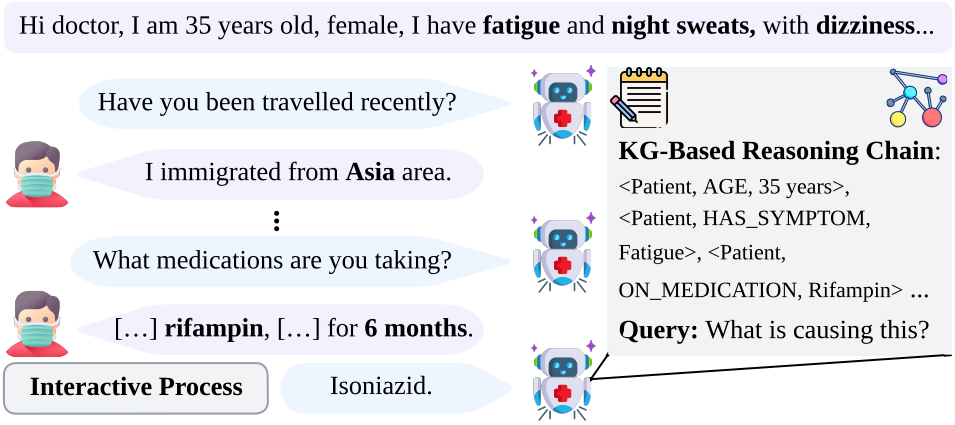

Figure 1: An illustration of an interactive medical QA, where clinical facts from patient responses are integrated into the doctor agent by knowledge triplets for multi-turn reasoning.

hop clinical reasoning.

In this work, we propose TriMediQ, a triplet-structured approach that uses a triplet agent to summarise patient agent responses into knowledge triplets and a trainable projection module that constructs a patient specific graph and injects it into a doctor agent to support structured multi-hop reasoning across dialogue turns. During the dialogue, the patient agent is solely responsible for answering questions posed by the doctor agent based on complete patient information. The doctor agent serves as a doctor who asks a question at each turn and performs diagnostic inference using the accumulated patient information. In parallel, the triplet agent continuously converts the patient response into knowledge triplets (Bodenreider, 2004) for the patient specific graph. To evaluate performance, we test TriMediQ on two interactive medical QA datasets (iCRAFT-MD and iMedQA (Johri et al., 2023, 2025)). Experimental results show that triplet-guided reasoning consistently outperforms five existing baselines, achieving up to a 10.4% absolute improvement on iMedQA. These findings demonstrate that structured triplet-based knowledge graphs enable more reliable multi-hop reasoning. Our main contributions are as follows:

- We propose **TriMediQ**, a triplet structured approach that converts patient information into a structured knowledge graph, enabling explicit modelling of clinical relations and multi-turn reasoning.
- We introduce a **trainable projection module** that constructs a patient-specific graph and injects graph-based representations into a doctor agent, facilitating structure-aware multi-hop reasoning.
- We conduct extensive experiments on two interactive medical QA datasets, demonstrating that TriMediQ consistently outperforms five strong baselines, achieving up to a **10.4% absolute improvement** in diagnostic accuracy.

2 Related Work

We review related research in the following three strands: medical QA systems that adapt LLMs to clinical domains, interactive agents designed for multi-turn reasoning, and KG-enhanced LLMs.

2.1 Medical Question Answering Systems

Medical QA systems have evolved from early rule-based approaches to advanced LLM-powered agents (Su et al., 2025; Almansoori et al., 2025). Benchmarks such as MultiMedQA (Singhal et al., 2023a) cover both multiple-choice and open-ended questions, while domain adaptation has been explored via biomedical corpora (Singhal et al., 2023b) and conversational datasets (Han et al., 2023). Beyond static QA, interactive frameworks are emerging. For example, MEDIQ (Li et al., 2024) simulates an doctor–patient dialogue setting where the doctor reasons over dialogue logs. In addition, ProMed (Ding et al., 2025) employs reinforcement learning with Shapley Information Gain rewards to encourage proactive

questioning. More recently, Med-U1 (Zhang et al., 2025) has explored large-scale RL to unify diverse medical reasoning tasks with a controllable reasoning length. Furthermore, DynamiCare (Shang et al., 2025) has modelled clinical decision-making as a dynamic multi-agent process that adapts specialist teams over multiple interaction rounds. However, the latter three approaches are not peer-reviewed and do not provide released code, which limits their reproducibility as baselines. Our approach extends interactive medical QA by incorporating structured knowledge representations to support multi-hop reasoning.

2.2 Interactive Models and Agents

Interactive dialogue agents extend standard QA systems by enabling iterative information acquisition and decision-making across multiple turns, as seen in domains such as negotiation, tutoring, and collaborative problem-solving (Abdelnabi et al., 2023; JU et al., 2024; Wang et al., 2025; Liu et al., 2025). Recent work has explored LLM-based agents for preference elicitation (Li et al., 2023; Andukuri et al., 2024) and multi-agent collaboration (Wu et al., 2023; Fang et al., 2025b), demonstrating their potential in complex interactive workflows. However, interactive modelling in the medical domain remains limited, with only a few approaches such as MEDIQ (Johri et al., 2023) exploring doctor–patient interaction settings. Existing methods primarily rely on unstructured dialogue histories, which makes it difficult to explicitly integrate clinically relevant information across turns. This limitation motivates the need for structured representations that support consistent multi-hop reasoning in interactive medical QA.

2.3 Knowledge Graph Integration Models

Several studies (Meng et al., 2024; Cao et al., 2024; Fang et al., 2025a) have shown that structured knowledge representations are beneficial for multi-hop QA. Graph-based reasoning has also been incorporated into conversational QA frameworks to integrate information from multiple sources (Wang et al., 2024; Jin et al., 2024). In addition, recent works (Ren et al., 2025; Kamalloo et al., 2023) have focused on the evaluation of open-domain and long-form QA with LLMs, revealing challenges in faithful assessment and reasoning consistency. QA datasets and resources have increasingly emphasised knowledge-intensive and domain-specific settings, including healthcare (Lunardi et al., 2025; Xu et al., 2023; Su et al., 2025). However, most KG-enhanced QA methods assume a static, offline knowledge source, disregarding the acquisition of incremental information during interactions. This motivates our approach, which dynamically constructs and updates structured knowledge representations to enable effective multi-hop reasoning in interactive medical QA.

3 Methodology

In the following sections, we present the task definition and the proposed TriMediQ. Our approach incorporates structured triplet knowledge into interactive medical QA by integrating a patient agent 3.3, a doctor agent 3.4, a triplet agent 3.5, and a graph-based projection module 3.6.

3.1 Task Definition

Interactive medical QA simulates an iterative consultation between a doctor agent and a patient agent (Bornstein and Emler, 2001; Masic, 2022). The doctor agent is initially given a multiple-choice question (MCQ), its candidate answers, and an initial patient description k_0. The patient agent has access to the complete patient record $\mathcal{K} = \{k_0, k_1, \ldots, k_n\}$ and reveals relevant subsets of $\mathcal{K}$ in response to the doctor's follow-up questions. At each turn t, the doctor agent either outputs a final answer or asks a follow-up question q_t. Given q_t, the patient agent returns a response $r_t \subseteq \mathcal{K}$, and the dialogue context is updated as $\mathcal{K}_t = \mathcal{K}_{t-1} \cup r_t$. The interaction terminates when the doctor agent decides on a final answer. The goal is to correctly answer the MCQ by integrating evidence accumulated over multiple turns.

3.2 Overview of TriMediQ

As shown in Fig. 2, TriMediQ is a triplet-structured approach for interactive medical QA. During the interaction, a **patient agent** responds to questions from a **doctor agent** based on complete patient information. Subsequently, the patient's responses are processed by a **triplet agent**, which converts free-form dialogue into knowledge triplets. The extracted

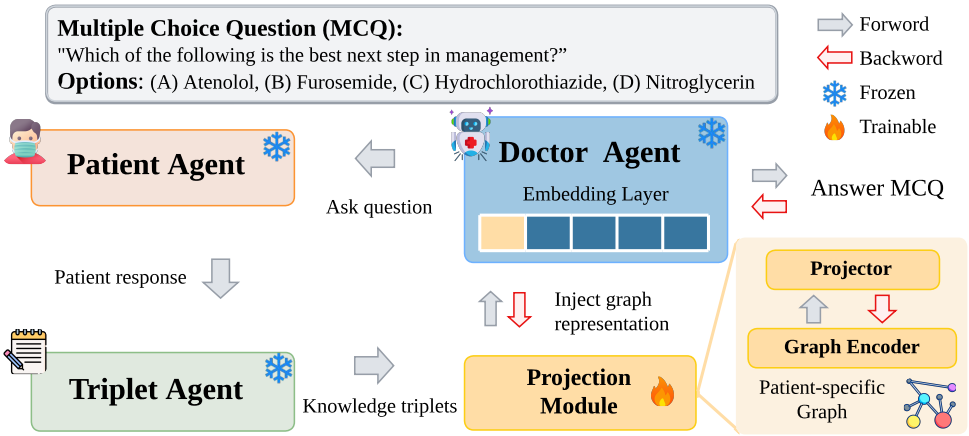

Figure 2: Overview of the TriMediQ training framework. Patient responses are parsed into structured knowledge triplets and organised as a patient-specific graph. The graph is encoded and injected into the embedding layer of the doctor agent via a trainable projection module. The doctor agent iteratively interacts with the patient to gather missing clinical evidence, while the injected graph representations enable structured multi-hop reasoning across dialogue turns.

triplets are passed into a trainable projection module. A graph encoder within this module incrementally constructs and encodes a patient-specific graph from the accumulated triplets, while a projector aligns the resulting graph representations with the hidden space of the doctor agent. The aligned representations are then used by the doctor agent to guide follow-up questioning and to produce the final result.

3.3 The Patient Agent

The patient agent is implemented as a frozen LLM that simulates a clinically realistic respondent in multi-turn diagnostic dialogues. At each turn t, given the doctor's follow-up question q_t and the complete patient record $\mathcal{K}$, the patient agent outputs a single sentence response $r_t \subseteq \mathcal{K}$ that is retrieved from the dataset associated with the patient record. The patient agent is not used for diagnostic reasoning or decision-making, and it does not introduce new clinical facts beyond those contained in the patient record.

Its output is strictly constrained to dataset-curated sentences. If no relevant sentence exists for the given query, the agent outputs a fixed refusal message: *"The patient cannot answer this question, please do not ask this question again"*, indicating that the question cannot be answered. By restricting responses to sentences within the dataset, the patient agent avoids speculative content and provides deterministic, well-defined inputs for subsequent triplet extraction. The prompt template for the patient agent is provided in Appendix A.2.

3.4 The Doctor Agent

The doctor agent is responsible for conducting the interactive consultation and predicting the final MCQ answer. It is implemented as a frozen LLM that explicitly integrates distributed patient information through a structured patient-specific graph, enabling reliable multi-hop reasoning across interaction turns. At each turn t, the doctor agent conditions on the target MCQ together with its candidate answers, and the patient-specific graph constructed from triplets accumulated up to turn $t-1$ via the projection module. Rather than reasoning over raw dialogue histories, the doctor agent operates on this structured representation, which encodes clinically relevant entities and relations collected across turns.

Based on the structured graph representation, the doctor agent performs multi-hop reasoning to decide between two actions: asking a targeted follow-up question to acquire missing clinical evidence, or outputting a final answer. As new patient responses are converted into triplets and incorporated into the patient-specific graph, the doctor agent incrementally updates its reasoning context, ensuring that both questioning and answer prediction remain grounded in structured and verifiable clinical evidence. (Prompt template in Appendix A.3)

3.5 The Triplet Agent

The triplet agent is responsible for converting patient responses into knowledge triplets. It is also a frozen LLM that takes as input the single-sentence response returned by the patient agent at turn t and extracts a set of knowledge triplets of the form $\langle \text{head}, \text{relation}, \text{tail} \rangle$. The triplet agent only performs information extraction. The output is restricted to entities and relations explicitly expressed in the input sentence, ensuring that the extracted triplets are directly grounded in patient information. The resulting triplets are accumulated across turns and passed to the projection module for subsequent graph construction and encoding. The prompt template for the triplet agent is provided in Appendix A.4.

3.6 The Projection Module

The projection module maps the patient-specific graph constructed from triplets into the representation space of the doctor agent, enabling structured clinical evidence to support multi-turn reasoning (Fig. 2). The module consists of two trainable components: a graph encoder and a projector. The graph encoder aggregates information from the accumulated triplets to produce a global graph representation, while the projector maps this representation into a prefix embedding compatible with the doctor agent.

Given the projected graph representation, the model is trained using a cross-entropy objective over the MCQ answer distribution. Each question contains four candidate options with a single correct answer, and is therefore formulated as a four-class classification problem. Let $\hat{\mathbf{y}} \in \mathbb{R}^4$ denote the predicted probability distribution over the four options, and $\mathbf{y}$ the one-hot ground-truth label. The training objective is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{4} y_c \log \hat{y}_c,$$

which directly supervises the alignment between the projected graph representation and the prediction space of the doctor LLM.

Graph Encoder: Triplets accumulated across interaction turns are assembled into a directed knowledge graph $\mathcal{G} = (V, E)$, where each node $v \in V$ represents a clinical entity and each edge $e = (u, r, v) \in E$ denotes a relation-labelled connection. Node and edge representations are initialised using Sentence-BERT embeddings (Reimers and Gurevych, 2019), computed from entity names and relation descriptions, respectively. We use a graph attention network (GAT) (Veličković et al., 2017) as the graph encoder to propagate information across the patient-specific graph. At layer ℓ, node representations are updated as follows:

$$\mathbf{H}^{(\ell+1)} = \text{GAT}\big(\mathbf{H}^{(\ell)}, \mathcal{G}\big),$$

where $\mathbf{H}^{(\ell)} \in \mathbb{R}^{|V| \times h}$ denotes the node embeddings at layer ℓ.

To obtain a global graph representation, we apply attention pooling over the node embeddings (Kim et al., 2025). A learnable attention mechanism assigns an importance weight to each node:

$$\alpha_v = \frac{\exp\big(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{H}_v)\big)}{\sum_{u \in V} \exp\big(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{H}_u)\big)},$$

and the graph-level representation is computed as a weighted sum:

$$\mathbf{h}_{\text{graph}} = \sum_{v \in V} \alpha_v \mathbf{H}_v.$$

Projector The projector maps the graph-level representation $\mathbf{h}_{\text{graph}}$ into the hidden space of the frozen doctor LLM. It is implemented as a lightweight multilayer perceptron with SiLU activation, followed by a linear projection and reshaping to form a fixed-length prefix embedding. This prefix is concatenated with the prompt embeddings at the embedding layer and passed to the doctor LLM, allowing structured graph information to inform both follow-up question generation and final prediction.

4 Experimental Setup

This section introduces the experimental setup, the research questions (RQ), the datasets and metrics used for evaluation.

All experiments follow the same interactive medical question answering protocol and share identical patient responses across all settings. The doctor agent is instantiated using open-source large language models of varying scales, including LLaMA-3 (8B and 70B), Qwen-2.5 (7B and 72B), and DeepSeek (14B), enabling

Table 1: Statistics of the evaluation datasets. iMedQA provides full train/dev/test splits, while iCRAFT-MD is used only for evaluation.

Dataset	Train	Dev	Test	Source
iMedQA	10,178	1,272	1,273	Derived from MedQA (Jin et al., 2021)
iCRAFT-MD	–	–	140	Dermatology (Johri et al., 2023, 2025)

Table 2: Accuracy (%) comparison across datasets with and without the proposed TriMediQ. Results are reported as mean ± standard deviation. * indicates a statistically significant improvement (paired t-test, $p < 0.05$).

LLM	iCRAFT-MD			iMedQA		
	Dialogue	TriMediQ	Δ	Dialogue	TriMediQ	Δ
LLaMA-3-8B	50.0 ± 4.2	$58.2 \pm 3.1^*$	+8.2	45.0 ± 1.4	$53.3 \pm 1.3^*$	+8.3
Qwen-2.5-7B	48.7 ± 4.6	$56.8 \pm 4.1^*$	+8.1	44.6 ± 1.5	$52.8 \pm 1.3^*$	+8.2
DeepSeek-14B	61.3 ± 3.6	$67.5 \pm 2.3^*$	+6.2	53.2 ± 1.7	$58.2 \pm 1.4^*$	+5.0
LLaMA-3-70B	72.1 ± 3.8	$76.3 \pm 2.9^*$	+4.2	60.9 ± 1.2	$71.3 \pm 1.2^*$	+10.4
Qwen-2.5-72B	71.4 ± 4.1	$77.1 \pm 3.8^*$	+5.7	61.5 ± 1.2	$71.4 \pm 1.3^*$	+9.9

evaluation across different backbone capacities. The patient agent and triplet extraction agent are fixed to Qwen-2.5-72B throughout all experiments to ensure consistent and high-quality patient responses and information extraction. Commercial models are not used as doctor agents, because they do not support the control-level interventions required by our experimental protocol. All backbone models remain frozen during training and evaluation.

Doctor Variants. We evaluate multiple configurations that differ only in how the information collected during interaction is provided to the doctor agent. Apart from this difference, the interaction protocol, stopping criteria, and answer selection procedure are kept identical to ensure a fair comparison.

4.1 Research Questions

Based on the above experimental design, our study aims to answer the following research questions:

- **RQ1:** How much does TriMediQ improve diagnostic accuracy over dialogue-only reasoning in interactive medical QA?
- **RQ2:** What is the effect of graph-based triplet integration compared with alternative triplet injection strategies?
- **RQ3:** How does imperfect triplet extraction affect the performance of TriMediQ?
- **RQ4:** How does TriMediQ affect diagnostic accuracy and interaction efficiency?

4.2 Datasets

iMedQA is reformulated from MedQA (Jin et al., 2021) into an interactive setting, where each case contains a complete patient record and an associated MCQ, with available training, development, and test splits. **iCRAFT-MD** is a dermatology-focused dataset comprising 140 cases (100 from an online question bank and 40 clinician-authored) (Johri et al., 2023, 2025), following the same interactive format but provided only as a test set without training or development splits. Table 1 summarises their main statistics.

4.3 Evaluation Metrics

Performance is measured using diagnostic accuracy and interaction efficiency, defined as the average number of questions asked before termination. Accuracy reflects the sufficiency of collected clinical information for correct diagnosis, while average turns capture the efficiency of information-seeking under identical interaction protocols. These metrics collectively form a concise yet comprehensive evaluation framework for interactive medical QA systems, encompassing both diagnostic accuracy and information retrieval efficiency.

5 Results

In the following, we present and discuss the results of our experiments. We structure the analysis around the four research questions defined in Section 4, showing how each experiment provides evidence to address them.

Table 3: Comparison of knowledge integration strategies and triplet quality on iCRAFT-MD and iMEDQA. Relative improvement (%) is computed relative to the Triplet-T variant.

Model	Triplet-T	Triplet-E	TriMediQ	Graph-D	Rel. (%)
iCRAFT-MD					
LLaMA-3-8B	53.9 ± 2.0	55.5 ± 2.2	**58.2 ± 3.1**	55.0 ± 3.2	+8.0
Qwen-2.5-7B	53.0 ± 3.0	54.3 ± 3.2	**56.8 ± 4.1**	53.6 ± 4.2	+7.2
DeepSeek-14B	61.3 ± 1.3	63.0 ± 1.8	**67.5 ± 2.3**	64.8 ± 2.5	+10.1
LLaMA-3-70B	72.1 ± 2.0	74.2 ± 2.3	**76.3 ± 2.5**	74.1 ± 3.1	+5.8
Qwen-2.5-72B	72.6 ± 1.4	74.6 ± 1.7	**77.1 ± 2.1**	74.3 ± 4.0	+6.2
iMEDQA					
LLaMA-3-8B	49.9 ± 1.4	51.0 ± 1.1	**53.3 ± 1.3**	50.2 ± 1.7	+6.8
Qwen-2.5-7B	48.7 ± 1.4	50.5 ± 1.2	**52.8 ± 1.3**	49.8 ± 1.9	+8.4
DeepSeek-14B	55.3 ± 1.5	56.0 ± 1.1	**58.2 ± 1.4**	55.8 ± 1.6	+5.2
LLaMA-3-70B	63.1 ± 1.2	66.0 ± 1.2	**71.3 ± 1.2**	67.8 ± 1.8	+13.0
Qwen-2.5-72B	64.6 ± 1.2	67.2 ± 1.1	**71.4 ± 1.3**	68.0 ± 1.5	+10.5

Table 4: Accuracy across different interaction budgets and overall interaction efficiency on iCRAFT-MD and iMedQA. Δ denotes the absolute accuracy gain. Avg. denotes the mean number of dialogue turns used during inference.

Turns	iCRAFT-MD			iMedQA		
	Dialogue	TriMediQ	Δ	Dialogue	TriMediQ	Δ
2	63.2 ± 3.5	68.4 ± 3.2	+5.2	52.1 ± 1.5	63.8 ± 1.4	+11.7
4	66.8 ± 3.6	73.5 ± 3.0	+6.7	55.4 ± 1.3	68.9 ± 1.3	+13.5
6	69.9 ± 3.7	**76.3 ± 2.9**	+6.4	58.3 ± 1.2	**71.3 ± 1.2**	+13.0
8	71.5 ± 3.8	76.1 ± 2.9	+4.6	59.8 ± 1.2	71.1 ± 1.2	+11.3
10	**72.1 ± 3.8**	76.2 ± 3.0	+3.9	**60.9 ± 1.2**	71.0 ± 1.3	+10.1
Avg.	**8.8**	**6.3**	–	**9.4**	**6.7**	–

5.1 RQ1: Effectiveness on Interactive Medical QA

A central question in interactive medical QA is whether replacing raw conversational text with structured, clinically grounded triplets enables more accurate and efficient diagnostic reasoning. As we argued in this paper, structured representations can make cross-turn dependencies explicit, reduce noise from irrelevant phrasing, and allow the doctor to reason over clinical facts with KG rather than implicitly reconstructing them from dialogue sentences. We evaluate the effectiveness of the proposed triplet-structured approach by comparing dialogue-based reasoning with structured triplet integration. In the Dialogue setting, the doctor agent reasons solely over the accumulated dialogue history, whereas in TriMediQ, the extracted triplets are explicitly aggregated, encoded, and injected as structured evidence across turns. As shown in Table 2, the triplet-structured integration consistently improves accuracy on both iCRAFT-MD and iMedQA across all LLM backbones. On iCRAFT-MD, gains include +8.2% for LLaMA-3-8B and +6.2% for DeepSeek-14B. On iMedQA, larger improvements are observed, reaching +10.4% for LLaMA-3-70B and +9.9% for Qwen-2.5-72B. In answer to RQ1, these results indicate that explicitly structuring and aggregating triplet-level evidence provides a more effective interface for integrating multi-turn clinical information than unstructured dialogue alone, enabling more reliable interactive medical QA.

5.2 RQ2: Effectiveness of Graph-Based Triplet Integration

A key question in triplet-based knowledge integration is whether simply appending extracted triplets to the doctor's prompt is sufficient, or whether a dedicated trainable projection mechanism provides additional benefits. To answer this, we compare different triplet injec-

tion strategies, including *Triplet-T*, which appends triplets directly to the prompt, *Triplet-E*, which projects pooled triplet representations, and TriMediQ, which injects graph-encoded triplets through a trainable projection module. As shown in Table 3, TriMediQ consistently outperforms both *Triplet-T* and *Triplet-E* across all backbone models and both datasets. For example, on iMedQA with LLaMA-3-70B, TriMediQ achieves 71.3, compared with 63.1 for *Triplet-T* and 66.0 for *Triplet-E*. Similar trends are observed on iCRAFT-MD, where graph-based integration yields the strongest performance across all evaluated backbones. In response to RQ2, the results show that the proposed graph-based projection provides a more effective way to inject structured patient knowledge into the frozen doctor agent than flat text-based or pooled triplet representations.

5.3 RQ3: The Effect of Imperfect Triplet Extraction

A further question is how imperfect triplet extraction affects the performance of TriMediQ. Since extracted triplets may be incomplete or noisy in realistic settings, it is important to assess whether the proposed graph-based integration remains useful when triplet quality is degraded. To examine this, we drop 20% of the extracted triplets at inference time, denoted as *Graph-D* in Table 3. As expected, performance decreases relative to the full TriMediQ setting on both datasets and across all backbones. However, *Graph-D* remains consistently competitive and generally outperforms the simpler *Triplet-T* baseline, while often remaining close to or above *Triplet-E*. For instance, on iMedQA with Qwen-2.5-72B, TriMediQ achieves 71.4, while *Graph-D* still reaches 68.0 despite degraded triplet coverage. In response to RQ3, these results suggest that the proposed TriMediQ approach is reasonably robust to imperfect triplet extraction and can preserve the benefit of structured knowledge integration.

5.4 RQ4: Interaction Efficiency Analysis

An important practical question in interactive medical QA is whether structured triplet integration enables the doctor agent to diagnose more accurately and efficiently. Thus, this analysis focuses on interaction efficiency, examining whether structured triplet integration enables the doctor agent to reach accurate diagnoses with fewer dialogue turns. LLaMA-3-70B serves as the backbone of the doctor agent. As shown in Table 4, TriMediQ consistently achieves higher accuracy than dialogue-only reasoning at the same number of turns on both iCRAFT-MD and iMedQA. Notably, TriMediQ reaches its peak performance earlier than dialogue-only reasoning. For example, on iCRAFT-MD, TriMediQ attains its peak accuracy at 6 turns, whereas dialogue-only reasoning continues to improve until 10 turns.

Beyond per-turn accuracy, TriMediQ reduces the number of interactions required during inference. As reported in the Avg. row of Table 4, the average number of dialogue turns decreases from 8.8 to 6.3 on iCRAFT-MD and from 9.4 to 6.7 on iMedQA. In answer to RQ4, these results indicate that structured triplet aggregation enables more efficient information usage, allowing accurate predictions to be reached earlier with fewer interactions.

6 Conclusions

Interactive medical QA is challenging because clinically relevant information is distributed across multiple dialogue turns and must be integrated for accurate diagnosis. Existing methods largely rely on dialogue histories, offering limited support for explicit multi-hop clinical reasoning. We proposed **TriMediQ**, a triplet-structured framework that converts patient responses into clinical triplets, aggregates them into a patient-specific knowledge graph, and injects structured evidence into a frozen doctor LLM via a lightweight projection module. Experiments on iCRAFT-MD and iMedQA show consistent improvements over dialogue-only baselines, achieving up to **10.4%** gains in diagnostic accuracy. These results indicate that triplet-based knowledge graphs provide an effective and practical interface for enhancing interactive medical QA and represent a practical step toward more reliable LLM-based clinical assistants. Finally, although our evaluation focuses on multiple-choice QA to reproducible benchmarking, real clinical reasoning often involves open-ended hypotheses. Future work will extend this framework to richer ontologies and clinician-authored dialogues.

Limitations

A primary limitation of this work is the reliance on interactive datasets that provide both rich patient records and turn-by-turn diagnostic dialogues. At present, such resources are limited to iMEDQA and iCRAFT-MD, constraining the diversity of conditions and specialities we can evaluate. Second, while our approach demonstrates consistent gains using knowledge triplets, the relation schema is fixed and may omit clinically relevant but less common relations, potentially limiting coverage in specialised domains. Third, our patient system is simulated and optimised for factuality. The real-world patients may provide incomplete, ambiguous, or noisy information that could challenge the triplet extraction. Finally, although our evaluation focuses on multiple-choice question answering to enable reproducible benchmarking, real clinical reasoning often involves open-ended hypotheses and longitudinal case management. Future work should expand the approach to broader datasets and explore deployment with real or semi-synthetic patient interactions in free-form.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2023. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. *Preprint*, arXiv:2309.17234.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. 2025. Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678*.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2024. Stargate: Teaching language models to ask clarifying questions. *Preprint*, arXiv:2403.19154.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Brian H Bornstein and A Christine Emler. 2001. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of evaluation in clinical practice*, 7(2):97–107.

Jiahang Cao, Jinyuan Fang, Zaiqiao Meng, and Shangsong Liang. 2024. Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Computing Surveys*, 56(6):1–42.

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*, 99(6):1120–1129.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2025a. Kirag: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. *Proceedings of the 63st Annual Meeting of the Association for Computational Linguistics*, 99(6):1120–1129.

Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, Zhaochun Ren, Nikos Aletras, Xi Wang, Han Zhou, and Zaiqiao Meng. 2025b. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems. *arXiv preprint arXiv:2508.07407*.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2024. Large language models on graphs: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering*, 99(6):1120–1129.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhou Ran Cai, and 1 others. 2025. An evaluation framework for conversational reasoning in clinical llms during patient interactions. *Nature medicine*, 31(1):77–86.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2023. Testing the limits of language models: A conversational framework for medical ai assessmentr. *medRxiv*, 99(6):1120–1129.

Da JU, Song Jiang, Andrew Cohen, Aaron Foss, Sasha Mitts, Arman Zharmagambetov, Brandon Amos, Xian Li, Justine T Kao, Maryam Fazel-Zarandi, and Yuandong Tian. 2024. To the globe (ttg): Towards language-driven guaranteed travel planning. *Preprint*, arXiv:2410.16456.

Ehsan Kamalloo, Charles LA Clarke, and Davood Rafiei. 2023. Limitations of open-domain question answering benchmarks for document-level reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2123–2128.

Jeongho Kim, Chanyeong Heo, and Jaehee Jung. 2025. Recdap: Relation-based conditional diffusion with attention pooling for few-shot knowledge graph completion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2848–2852.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*.

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*, pages 1120–1129.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Siwei Liu, Jinyuan Fang, Han Zhou, Yingxu Wang, and Zaiqiao Meng. 2025. Sew: Self-evolving agentic workflows for automated code generation. *arXiv preprint arXiv:2505.18646*.

Riccardo Lunardi, Michael Soprano, Paolo Coppola, Vincenzo Della Mea, Stefano Mizzaro, and Kevin Roitero. 2025. Pils of knowledge: A synthetic benchmark for evaluating question answering systems in healthcare. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3648–3658.

Izet Masic. 2022. Medical decision making-an overview. *Acta Informatica Medica*, 30(3):230.

Zhaohan Meng, Siwei Liu, Shangsong Liang, Bhautesh Jani, and Zaiqiao Meng. 2024. Heterogeneous biomedical entity representation learning for gene–disease association prediction. *Briefings in Bioinformatics*, 25(5):bbae380.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tianyu Ren, Zhaoyu Zhang, Hui Wang, and Karen Rafferty. 2025. Obella: Open the book for evaluating long-form large language model answers in open-domain question answering. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1109–1119.

Tianqi Shang, Weiqing He, Charles Zheng, Lingyao Li, Li Shen, and Bingxin Zhao. 2025. Dynamicare: A dynamic multi-agent framework for interactive and open-ended medical decision-making. *arXiv preprint arXiv:2507.02616*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023b. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2025. Kgarevion: an ai agent for knowledge-intensive biomedical qa. In *The Thirteenth International Conference on Learning Representations*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv*.

Duokang Wang, Linmei Hu, Rui Hao, Yingxia Shao, Xin Lv, Liqiang Nie, and Juanzi Li. 2024. Let me show you step by step: An interpretable graph routing network for knowledge-based visual question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1984–1994.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023. Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems*, 36:30840–30861.

Yingxu Wang, Siwei Liu, Jinyuan Fang, and Zaiqiao Meng. 2025. Evoagentx: An automated framework for evolving agentic workflows. *arXiv preprint arXiv:2507.03616*.

Eyal Winter. 2002. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, Wei Jin, Joyce Ho, and Carl Yang. 2023. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. *Proceedings of the 62st Annual Meeting of the Association for Computational Linguistics*.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. 2025. Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning. *arXiv preprint arXiv:2506.12307*.

A Appendix

A.1 Computing Hardware

All experiments were conducted on GPU clusters equipped with NVIDIA GH200 GPUs. These resources provide sufficient memory and computational capacity to support large-scale model inference and training under the experimental settings considered in this work.

A.2 Patient Agent Prompts

Atomic fact decomposition Prompt

You are a truthful medical assistant that understands the patient's information.
Break the following patient information into a list of independent atomic facts, with one piece of information in each statement. Each fact should only include the smallest unit of information, but should be self-contained.
"She says that, despite going to bed early at night, she is unable to fall asleep. She denies feeling anxious or having disturbing thoughts while in bed. Even when she manages to fall asleep, she wakes up early in the morning and is unable to fall back asleep. She says she has grown increasingly irritable and feels increasingly hopeless, and her concentration and interest

Atomic fact decomposition Prompt (cont'd)

at work have diminished. The patient denies thoughts of suicide or death. Because of her diminished appetite, she has lost 4 kg (8.8 lb) in the last few weeks and has started drinking a glass of wine every night instead of eating dinner. She has no significant past medical history and is not on any medications."
Response with the list of atomic facts and nothing else, prepend each fact by an index starting from 1. No sub-list allowed.

Example LLM Output

1.patient goes to bed early at night but is unable to fall asleep.
2.patient denies feeling anxious or having disturbing thoughts while in bed.
3.patient wakes up early in the morning and is unable to fall back asleep.
4.patient has grown increasingly irritable and feels increasingly hopeless.
5.patient's concentration and interest at work have diminished.
6.patient denies thoughts of suicide or death.
7.patient has lost 4 kg (8.8 lb) in the last few weeks.
8.patient started drinking a glass of wine every night instead of eating dinner.
9.patient has no significant past medical history.
10.patient is not on any medications."

patient Response Generation Prompt

You are a truthful assistant that understands the patient's information, and you are trying to answer questions from a medical doctor about the patient.
Below is a list of factual statements about the patient:
1.patient goes to bed early at night but is unable to fall asleep.
2.patient denies feeling anxious or having disturbing thoughts while in bed.
3.patient wakes up early in the morning and is unable to fall back asleep.
4.patient has grown increasingly irritable and feels increasingly hopeless.
5.patient's concentration and interest at work have diminished.
6.patient denies thoughts of suicide or death.
7.patient has lost 4 kg (8.8 lb) in the last few weeks.
8.patient started drinking a glass of wine every night instead of eating dinner.
9.patient has no significant past medical history.
10.patient is not on any medications.
Question from the doctor: "What time do you usually go to bed at night?"
Which of the above atomic factual statements answer the question? Select at most two statements. If no statement answers the question, simply say "The patient cannot answer this question, please do not ask this question again." Answer only what the question asks for. Do not provide any analysis, inference, or implications. Respond by selecting all statements that answer the question from above ONLY and NOTHING ELSE.

Example LLM Output

patient goes to bed early at night but is unable to fall asleep.

A.3 Doctor Agent Prompts

TriMediQ-doctor — Structured Assessment Prompt

You are a medical doctor answering real-world medical entrance exam questions. Based on your understanding of basic and clinical science, medical knowledge, and mechanisms underlying health, disease, patient care, and modes of therapy, answer the following multiple choice question. Select one correct answer from A to D. Base your answer on current standards and evidence-based guidelines.
Task: You will be asked to reason through the current patient's information using both the patient initial information and structured clinical facts (UMLS-style triplets).
patient initial information: "A 40-year-old woman presents with difficulty falling asleep, diminished appetite, and tiredness for the past 6 weeks."
Knowledge triplets:
(patient, HAS_SYMPTOM, insomnia)
(patient, HAS_SYMPTOM, reduced appetite)
(patient, HAS_SYMPTOM, fatigue)
(insomnia, DURATION, six weeks)

TriMediQ-doctor — Structured Assessment Prompt (cont'd)

Given the information above, your task is to choose one of four options that best answers the inquiry.
INQUIRY: "Which of the following is the best course of treatment in this patient?"
OPTIONS: A: Diazepam, B: Paroxetine, C: Zolpidem, D: Trazodone
Medical conditions are complex, so you should consider the relevant specialty, enumerate the key features required for a correct judgement, and assess whether the structured evidence provided by the triplets is sufficient. Think step by step about the patient initial information, the triplets, the inquiry, and the options. **Do not provide the answer choice**; keep your response to one paragraph.

Example LLM Output – Case 1: Follow-up Question

Have you also been experiencing feelings of sadness or loss of interest in daily activities?

Example LLM Output – Case 2: Prediction

B

A.4 Triplet Agent Prompts

TriMediQ Triplet Extraction – Initial Information Prompt

You are a medical terminology normaliser and knowledge graph constructor. Your task is to extract structured, clinically relevant facts from the patient's initial presentation, representing each fact as a UMLS-style triplet in the format (subject, relation, object). Only include directly observed and verifiable information without inferring diagnoses, causes, or risks not explicitly stated. Use only medically meaningful relations such as HAS_SYMPTOM, HAS_SIGN, HAS_MEDICATION, HAS_ALLERGY, HAS_PAST_MEDICAL_HISTORY, FAMILY_HISTORY_OF, SOCIAL_HISTORY_OF, FINDING_OF, TEMPORAL_PATTERN, SEVERITY, LOCATION, LATERALITY, DURATION, COURSE, and FUNCTIONAL_IMPACT. When appropriate, express information in a two-step form by stating a base fact and then attaching qualifiers to that concept, avoiding compounded objects. Extract at most three triplets per sentence, outputting NO_TRIPLET if no relevant fact is present, and maintain consistent clinical terminology across extractions.

patient initial information: "A 40-year-old woman presents with difficulty falling asleep, diminished appetite, and tiredness for the past 6 weeks."

Example LLM Output

(patient, HAS_SYMPTOM, insomnia)
(insomnia, DURATION, six weeks)
(patient, HAS_SYMPTOM, reduced appetite)
(patient, HAS_SYMPTOM, fatigue)

TriMediQ Triplet Extraction – Incremental Update Prompt

You are a medical terminology normaliser and knowledge graph constructor. Given the patient's existing knowledge triplets and new conversational information, extract up to three additional UMLS-style triplets in the format (subject, relation, object). Only include directly observed and verifiable information from the new input without inferring diagnoses, causes, or risks not explicitly stated. Avoid duplicating facts already present in the existing triplets. Use only medically meaningful relations such as HAS_SYMPTOM, HAS_SIGN, HAS_MEDICATION, HAS_ALLERGY, HAS_PAST_MEDICAL_HISTORY, FAMILY_HISTORY_OF, SOCIAL_HISTORY_OF, FINDING_OF, TEMPORAL_PATTERN, SEVERITY, LOCATION, LATERALITY, DURATION, COURSE, and FUNCTIONAL_IMPACT. When appropriate, express information in a two-step form by stating a base fact and then attaching qualifiers to that concept, avoiding compounded objects. Extract at most three triplets per sentence, outputting NO_TRIPLET if no relevant fact is present, and ensure consistent terminology with the existing knowledge graph.

Existing triplets: (patient, HAS_SYMPTOM, insomnia)
(patient, HAS_SYMPTOM, reduced appetite)
(patient, HAS_SYMPTOM, fatigue)

New patient information: "Her sleep problem is worse in the early mornings and she often needs daytime naps. She drinks two cups of coffee after dinner."

Example LLM Output

(insomnia, TEMPORAL_PATTERN, worse in early mornings)
(patient, FUNCTIONAL_IMPACT, daytime naps)
(patient, SOCIAL_HISTORY_OF, evening caffeine use)