Triplet-Structured Knowledge Integration for Multi-Turn Medical Reasoning

Zhaohan Meng[1], Zaiqiao Meng[1⋆], Siwei Liu[2], and Iadh Ounis[1]

[1] School of Computing Science, University of Glasgow, UK
[2] School of Natural and Computing Science, University of Aberdeen, UK

Abstract. Large Language Models (LLMs) have shown strong performance on static medical Question Answering (QA) tasks, yet their reasoning often deteriorates in multi-turn clinical dialogues where patient information is scattered across turns. This paper introduces TriMediQ, a triplet-structured approach that enhances the reasoning reliability of LLMs through explicit knowledge integration. TriMediQ first employs a frozen triplet extraction LLM to convert patient responses into clinically grounded triplets, ensuring factual precision via constrained prompting. These triplets are incorporated into a patient-specific Knowledge Graph (KG), from which a trainable projection module consisting of a graph encoder and a projector captures relational dependencies while keeping all LLM parameters frozen. During inference, the projection module guides multi-hop reasoning over the KG, enabling coherent clinical dialogue understanding. Experiments on two interactive medical QA benchmarks show that TriMediQ achieves up to 10.4% improvement in accuracy over five existing baselines on the iMedQA dataset. These results demonstrate that structuring patient information as triplets can effectively improve the reasoning capability of LLMs in multi-turn medical QA.

1 Introduction

Large Language Models (LLMs) have achieved remarkable results on static and single-turn medical Question Answering (QA) benchmarks such as MedQA [16] and PubMedQA [17]. However, these settings fail to capture the interactive nature of real-world clinical scenarios, where diagnoses are formed through iterative questioning [24]. In practice, clinicians indeed refine hypotheses over multiple turns, selectively eliciting missing clinical facts and incrementally connecting them into a coherent Knowledge Graph (KG) that supports the final decision [10]. Therefore, static and single-turn QA evaluations typically overestimate the reliability of LLMs in interactive contexts and thereby encourage premature answers to questions when information is incomplete [27,20]. A central requirement in interactive medical QA is the ability to explicitly connect distributed clinical facts across dialogue turns in order to perform reliable multi-hop reasoning [21]. For instance, an expert LLM system might need to connect a

⋆ Corresponding author

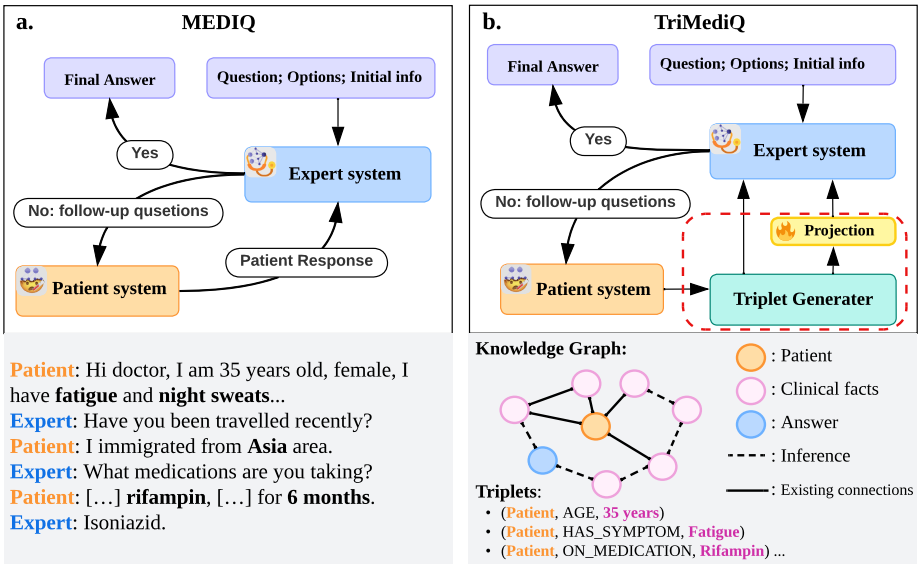

Fig. 1. Interactive Medical Question Answering: a. The **MEDIQ** framework requires the expert to identify key clinical facts (e.g., **fatigue**, **night sweats**, **rifampin**) directly from dialogue. b. The proposed **TriMediQ** highlights these clinical facts as triplet-based KG, where pink nodes denote different clinical facts in the KG.

symptom mentioned early in the conversation with a medication revealed later in the dialogue. This motivates the development of interactive QA paradigms where LLMs must actively engage in multi-turn information gathering and structured reasoning rather than passively responding to isolated questions.

Interactive dialogue medical QA systems, such as MEDIQ [21], have taken an important first step towards modelling multi-turn clinical reasoning [34,15]. As illustrated in Figure 1, MEDIQ uses two LLMs: a *patient* system, which has full access to the medical record and simulates a patient by returning context-grounded answers, and an *expert* system, which must decide at each turn whether to predict the final answer or to ask a follow-up question. Although human clinicians can internally distil and organise important facts from patient conversations, current LLMs tend to process the history of the dialogue as a flat sequence of tokens [12]. This limits their ability to integrate and retrieve relevant information across multiple turns [23]. For example, the MEDIQ-expert LLM system reasons over raw dialogue logs, where clinical facts appear in different sentences without explicit connections, thus resulting in inaccurate predictions or follow-up questions that deviate from the subject of Multiple Choice Questions (MCQ). This forces the expert LLM to implicitly recover clinical fact connections from dialogue logs, which is often unreliable when information needs to be integrated across multiple turns. To address this limitation, we propose to instead inte-

grate these distributed clinical facts into a triplet-structured knowledge graph, enabling the expert system to perform more reliable multi-turn reasoning.

In this work, we propose TriMediQ, which is an approach that extends the MEDIQ framework while retaining the design of the patient and expert systems [21]. The patient system receives the complete patient record along with the current follow-up question, and returns a factual response by selecting only the directly relevant atomic facts from the record. The key difference is that TriMediQ provides the expert system with a triplet-structured KG constructed from UMLS-style triplets [2] such as (*Patient, Has_Symptom, Fatigue*). These triplets are incrementally accumulated during the interaction and organised into a semantic KG that captures both the factual content and the relationships among entities. Since effective reasoning requires understanding not only individual facts but also their interconnections, we employ a trainable graph encoder to capture relational structures and a projection module that maps the graph into a prefix representation for the expert, thereby enabling multi-hop reasoning.

TriMediQ operates in two steps while keeping all backbone LLMs frozen. In the fine-tuning stage, the projection module is trained on the iMedQA dataset. In the inference stage, we evaluate TriMediQ on two interactive benchmarks (iCRAFT-MD and iMedQA [18,19] test datasets), where triplet-guided reasoning yields improvements over five baselines, with gains of up to 10.4% improvement on iMedQA. These results suggest that transforming dialogue into a triplet-based KG enables the expert system to perform more reliable multi-hop reasoning and achieve higher predictive accuracy while remaining compatible with general-purpose LLMs. Our main contributions are as follows:

1. We propose **TriMediQ**, a triplet-structured approach that incrementally converts patient information into a triplet-based KG, enabling explicit clinical connections and multi-turn reasoning.
2. We develop a **projection module** that integrates triplet-based graph representations with the expert LLM, supporting structured multi-hop reasoning and improving predictive accuracy.
3. We demonstrate that TriMediQ achieves substantial gains in interactive medical QA benchmarks, with **up to a 10.4% absolute improvement** in diagnostic accuracy over five baselines.

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.

Medical Question Answering Systems: Medical QA systems have evolved from early rule-based approaches to advanced LLM-powered agents [32,1]. Benchmarks such as MultiMedQA [30] cover both multiple-choice and open-ended questions, while domain adaptation has been explored via biomedical corpora [31] and conversational datasets [8]. Beyond static QA, interactive frameworks are

emerging. For example, MEDIQ [21] simulates an expert–patient dialogue setting where the expert reasons over dialogue logs. In addition, ProMed [6] employs reinforcement learning with Shapley Information Gain rewards to encourage proactive questioning. More recently, Med-U1 [37] has explored large-scale RL to unify diverse medical reasoning tasks with a controllable reasoning length. Furthermore, DynamiCare [29] 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.

Knowledge Graph Integration Models: Incorporating structured KG into LLMs has emerged as an effective strategy to improve reasoning, accuracy and interpretability [36,4]. Early approaches augmented neural architectures with representations such as KGs, leveraging graph embeddings to inject relational structures into downstream tasks [35,22]. More recent work integrated KG with LLM through retrieval-augmented generation [7] and prompt engineering [14] to access and reason over explicit entity and relation structures. In the biomedical domain, KG-enhanced models have been used for interaction prediction [26] and medical QA [25,32], demonstrating improved performance in knowledge-intensive settings. However, most methods assume an offline static KG, which limits their applicability in interactive scenarios where information is revealed incrementally [11]. Our TRIMediQ approach differs by dynamically constructing a patient-specific KG during the dialogue, which combines the adaptability of interactive agents with the multi-hop reasoning benefits of KG integration.

3 Methodology

In the following sections, we describe the task definition and the proposed TriMediQ approach. Our approach introduces structured triplet knowledge into interactive medical QA by combining a frozen patient LLM, a frozen expert LLM, a triplet generator, and a graph-based projection module.

Task Definition: The interactive medical QA task simulates an iterative consultation between *an expert* system and *a patient system* [3,24]. The expert initially receives the target multiple-choice question along with its answer options as well as the patient's initial description, k_0, containing basic demographics and presenting complaints. The patient system initially contains the complete patient record $\mathcal{K} = k_0, k_1, \ldots, k_n$, and responds to the expert system's question based on the patient record corresponding to that MCQ. At turn t, the expert is available to provide a final answer or ask follow-up questions q_t based on the initial input and accumulated knowledge. Given q_t and $\mathcal{K}$, the patient responds with $r_t \subseteq \mathcal{K}$, after which the expert updates its knowledge as $\mathcal{K}t = \mathcal{K}_{t-1} \cup r_t$. The goal is to build an effective *expert system* capable of identifying and collecting the necessary subset $\mathcal{K}^* \subseteq \mathcal{K}$ through targeted questions, enabling the expert

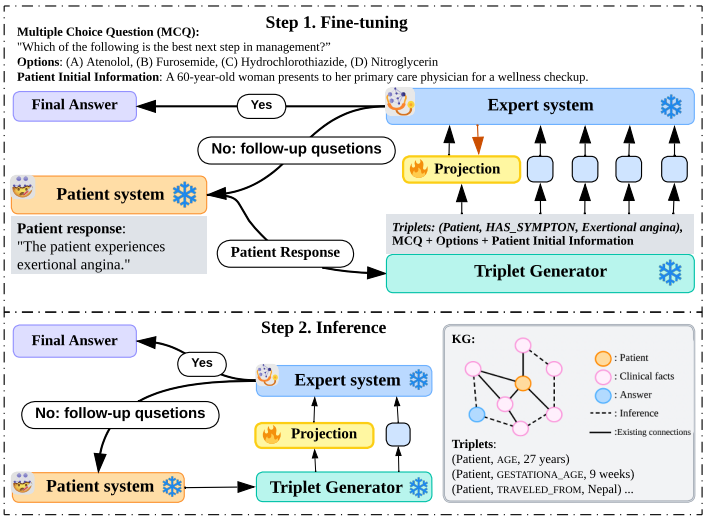

Fig. 2. Overview of the TriMediQ for Interactive Medical QA: The framework consists of two stages: (i) fine-tuning a projection module with a structured triplet graph; (ii) using the trained projection to support inference during interactive QA. The **Triplet Generator** converts patient responses into UMLS-style triplets, while the **Projection Module** (graph encoder + projector) encodes these triplets and injects them into the frozen expert LLM via prefix tuning.

to accurately answer the MCQ, despite relevant facts being sparsely distributed across turns and connected only by weak semantic links.

Overview of TriMediQ: TriMediQ is a triplet-structured approach that integrates structured clinical knowledge into interactive medical QA. As shown in Figure 2, it operates in two stages. In the first stage, a projection module is fine-tuned to align the triplet-based graph encoder with the frozen expert LLM. This design ensures that the expert can subsequently reason over KG-derived representations without updating its own parameters, thereby mitigating overfitting and reducing computational cost. In the second stage, the trained projection is applied during multi-turn diagnostic dialogues, where newly extracted triplets are incrementally accumulated into the KG and injected into the expert to enable multi-hop reasoning. The following sections describe each component in detail: the Patient LLM System (Sec. 3.1), the Expert LLM System (Sec. 3.2), the Triplet Generator (Sec. 3.3), and the Projection Module (Sec. 3.4).

3.1 The Patient System

To isolate the contribution of the expert system and ensure high-quality inputs to the projection module for triplet extraction, we employ a fixed patient system shared across TriMediQ and all MEDIQ baselines. This system simulates a

clinically realistic patient in multi-turn diagnostic dialogues. At each turn t_s, it receives the complete patient record $\mathcal{K}$ and the expert's current follow-up question q_{t_s}, and returns a set of atomic facts $r_t \subseteq \mathcal{K}$ that directly answer q_{t_s}. If no fact in the record can address the query, the system outputs a fixed refusal message: *"The patient cannot answer this question, please do not ask this question again"*. By constraining responses to atomic facts explicitly stored in the record, the patient system ensures high factuality (faithfulness to the patient's record) and relevance (directly addressing the expert's question). This design yields high precision for triplet extraction, reduces noise in the KG, and prevents conversational drift, allowing each extracted triplet to be clearly attributed to verifiable information in the record.

3.2 The Expert System

The expert system in TriMediQ explicitly integrates distributed clinical facts across dialogue turns, enabling reliable multi-hop reasoning for diagnostic decision-making. At each turn t, the expert receives three inputs: (i) the accumulated triplets $\mathcal{T}_{t-1}$ from previous turns, organised as a knowledge graph; (ii) the current MCQ with candidate answers; and (iii) a decision rule for whether to issue a final prediction or request additional information. Rather than concatenating dialogue logs as a flat sequence, the expert reasons over the evolving graph, connecting entities and relations across turns to build a coherent representation of the patient's condition. Based on this structured view, the expert performs multi-hop inference and chooses between two actions: outputting a diagnosis with confidence scores or generating a targeted follow-up query to close specific knowledge gaps. The knowledge graph is expanded whenever new patient responses are converted into triplets, ensuring that the expert's reasoning remains grounded in structured and verifiable evidence. The interaction process continues until the confidence in a prediction surpasses a predefined threshold or a maximum number of turns is reached.

3.3 Triplet Generator

The Triplet Generator extracts structured clinically grounded knowledge from patient responses, corresponding to the triplet extraction step in Figure 2. At each interaction step, it receives the current patient utterance together with previously extracted triplets, and generates new ones exclusively from the latest response. Each triplet adheres to a controlled schema of medically meaningful relations (e.g., *Has_Symptom*, *History_Of*, *Duration*), ensuring compatibility with UMLS biomedical ontologies. To maintain consistency and avoid redundancy, newly generated triplets are validated against previously extracted ones before being added to the knowledge graph.

The patient system produces atomic fact sentences, which typically contain only one or two clinical facts. For example, from the response *"The patient has fatigue and night sweats"*, the system extracts triplets such as (*Patient,*

Has_ Symptom, Fatigue) and (*Patient, Has_ Symptom, Night_ sweats*). To reduce hallucination and improve reliability, the generation process is constrained to output only atomic verifiable relations. Base observations are introduced first, and then compositional modifiers such as severity or temporal pattern are added. To balance coverage with precision, the number of triplets is capped at three per turn, which prioritises clinically salient facts while suppressing spurious or noisy relations. These high-precision triplets provide a structured intermediate layer that bridges raw dialogue logs with the expert's reasoning process, supporting downstream projection, graph encoding, and multi-hop diagnostic decision-making within TriMediQ.

3.4 Projection Module

The projection module bridges the triplet-based knowledge graph with the frozen expert LLM via prefix tuning, as illustrated in Figure 2. It consists of two trainable components: a graph encoder and a projector. During fine-tuning, these modules are optimised with a cross-entropy objective over MCQ answer distributions. We adopt cross-entropy because the task is naturally formulated as multi-class classification, where each MCQ has a single correct answer. This objective provides direct supervision by minimising the divergence between predicted and gold answer distributions, ensuring stable optimisation and aligning the projected graph representation with the expert's prediction space. All parameters of the expert LLM remain frozen.

Graph Encoder: Triplets accumulated over turns are assembled into a directed graph $\mathcal{G} = (V, E)$, where nodes represent clinical entities, and edges denote relation-labelled links. The node and edge features are initialised with Sentence-Bert [28], which provides semantically meaningful embeddings of clinical entities and relations. SentenceBert has been widely shown to capture sentence-level semantics more effectively than static embeddings (e.g. Word2Vec [5]), thereby providing informative initial representations for graph-based reasoning. A graph neural network backbone [9] (e.g. GCN [13] or GAT [33]) performs message passing to generate contextualised node embeddings $\mathbf{H} \in \mathbb{R}^{|V| \times h}$. Mean pooling over $\mathbf{H}$ yields a structure-aware graph vector:

$$\mathbf{h}_{\text{graph}} = \frac{1}{|V|} \sum_{v \in V} \mathbf{H}_v.$$

Projector: A multilayer perceptron first expands $\mathbf{h}_{\text{graph}}$ to a hidden layer with SiLU activation, then projects this representation to match the expert's hidden size d. The output is reshaped into a fixed-length prefix embedding $P \in \mathbb{R}^{k \times d}$, where k is the prefix length. This prefix is concatenated with the MCQ prompt embeddings at the embedding layer and passed to the frozen expert LLM.

4 Experimental Setup

This section introduces the research questions that guide our study, the datasets used for evaluation, and the experimental setup used in our experiments.

4.1 Research Questions

We conduct experiments to evaluate the effectiveness of the TriMediQ approach in interactive medical QA. As mentioned in Section 3, in all experiments, the patient system and the expert LLM remain frozen. The variation across configurations lies in how the triplet knowledge produced by the frozen generator is processed and injected into the expert. We compare three configurations:

1. **Instruction Prompt (IP):** The frozen triplet generator produces UMLS triplets, which are appended verbatim to the MCQ prompt (question, options and initial information) without LLM gradient update.
2. **Projector Training (PT):** The generator outputs are pooled into a fixed-size vector and passed through a trainable projection layer that maps them into prefix embeddings, which are additional embedding vectors prepended to the expert's input sequence.
3. **Projection Fine-tuning (PGT):** The generator outputs are first converted into a KG and encoded by a GCN-based graph encoder described in Section 3.4, which injects structured knowledge into the expert via prefix embeddings (see Figure 2).

The above three configurations allow us to isolate the effects of (i) using unstructured triplet text, (ii) learning a direct mapping from triplets to the expert's latent space, and (iii) leveraging graph-structured encoding before projection, allowing a fair comparison of different knowledge integration strategies. Our experiments aim to answer the following 4 research questions:

- **RQ1:** Does the proposed TriMediQ method outperform the MEDIQ baseline in interactive medical QA?
- **RQ2:** What is the performance difference between the *Instruction Prompt* and *Projection Fine-tuning* configurations when integrating triplet-based knowledge graph?
- **RQ3:** What is the performance difference between *Projector Training* and *Projection Fine-tuning* in terms of diagnostic accuracy?
- **RQ4:** How does diagnostic accuracy vary with dialogue depth when comparing MEDIQ and TriMediQ?

4.2 Datasets

We evaluate TriMediQ on two interactive benchmarks, iMedQA [16] and iCRAFT-MD [18,19], both released with the MEDIQ framework [21]. These datasets provide patient–expert interaction settings without requiring re-conversion from static sources. Below, we provide a brief description of the datasets. Table 1 presents their main statistics.

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	Domain / Source
iMedQA	10,178	1,272	1,273	Derived from MedQA [16]
iCRAFT-MD	–	–	140	Dermatology; 100 from online bank, 40 clinician-authored [18,19]

iMedQA: This dataset has been reformulated from MedQA [16] into an interactive format, where each medical case contains a complete patient record and an associated MCQ. The dataset provides training, development, and test splits.
iCRAFT-MD: This is a dermatology-focused dataset comprising 140 cases, of which 100 are sourced from an online question bank and 40 are authored by clinicians [18,19]. It follows the same interactive format as iMedQA but is provided only as a test set without training or development splits.

5 Results

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

5.1 RQ1: Model Effectiveness on the Interactive Medical QA Task.

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 expert to reason over clinical facts with KG rather than implicitly reconstructing them from dialogue sentences. As shown in Table 2, our proposed approach consistently surpasses MEDIQ on both iCRAFT-MD and iMEDQA. On iCRAFT-MD, the TriMediQ yields notable gains, such as +8.2% for Llama-3-8B. On iMEDQA, improvements are even larger, with Llama-3-70B achieving around +10% absolute accuracy. These consistent gains across two datasets and different LLM sizes suggest that graph-level aggregation exposes multi-hop clinical dependencies that are difficult to recover from the conversation logs, and that structured projection provides a solution to integrate this evidence into frozen LLMs. In answer to RQ1, the results confirm that TriMediQ consistently outperforms MEDIQ on datasets and model scales, demonstrating its effectiveness for interactive medical QA.

5.2 RQ2: Effect of Projection Fine-tuning vs. Instruction Prompt

An important question in integrating triplet-based evidence is whether simply appending extracted triplets to the expert's prompt is sufficient, or whether

Table 2. Accuracy (%) comparison between **MEDIQ** and **TriMediQ** across datasets. Results are mean ± std, with best values **bold**.

LLM	iCRAFT-MD			iMEDQA		
	MEDIQ	TriMediQ	Gain	MEDIQ	TriMediQ	Gain
Llama-3-8B	50.0 ± 4.2	**58.2 ± 3.1**	↑8.2	45.0 ± 1.4	**53.3 ± 1.5**	↑8.3
Qwen-2.5-7B	48.7 ± 4.6	**56.8 ± 4.1**	↑8.1	44.6 ± 1.5	**52.8 ± 1.7**	↑8.2
DeepSeek-14B	61.3 ± 3.6	**67.5 ± 2.3**	↑6.2	53.2 ± 1.7	**58.2 ± 1.4**	↑5.0
Llama-3-70B	72.1 ± 3.8	**76.3 ± 2.9**	↑4.2	60.9 ± 1.2	**71.3 ± 1.6**	↑10.4
Qwen-2.5-72B	71.4 ± 4.1	**77.1 ± 3.8**	↑5.7	61.5 ± 1.2	**71.4 ± 1.3**	↑9.9

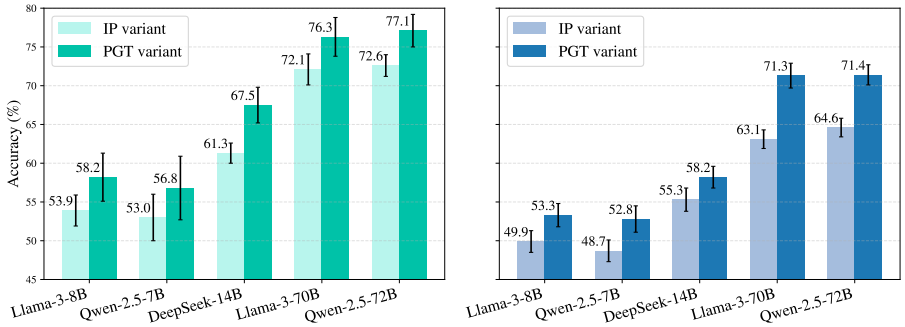

Fig. 3. Accuracy (%) comparison between IP and PGT across multiple LLMs on iCRAFT-MD (left) and iMEDQA (right).

fine-tuning a dedicated projection mechanism yields additional benefits. In the IP variant (Sec. 4), UMLS-style triplets are appended verbatim to the MCQ prompt without training, exposing the expert to all triplets' content but relying entirely on the LLM to interpret and integrate these facts in context. In contrast, the PGT variant accumulates triplets in a KG, encodes them with a GCN, and projects the structure-aware representation into a prefix via a trainable module (Sec. 3.4), thereby aligning the graph embedding space with the expert's internal task representation. As shown in Figure 3, PGT consistently outperforms IP across both datasets and LLM sizes. For example, Llama-3-8B improves from 53.9 % to 58.2 % in iCRAFT-MD and Qwen-2.5-72B improves from 64.6 % to 71.4 % in iMEDQA. These improvements indicate that fine-tuning the projection provides a stronger inductive bias than zero-shot in-context explanation, enabling the expert to better retrieve and integrate distributed evidence across turns and thereby improving diagnostic accuracy. Regarding RQ2, our experiments show that PGT yields clear gains over IP, showing that learning a structured mapping is more effective than zero-shot prompting.

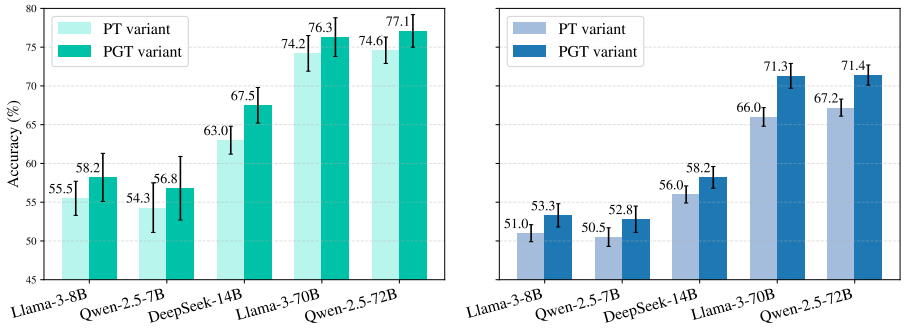

Fig. 4. Accuracy (%) comparison between PT and PGT across multiple LLMs on iCRAFT-MD (left) and iMEDQA (right).

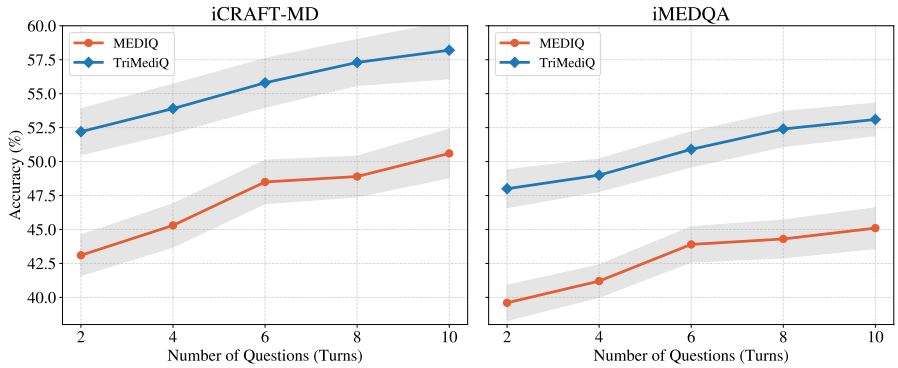

Fig. 5. Accuracy (%) as a function of dialogue turns. TriMediQ outperforms MEDIQ with consistent gains as the number of turns increases.

5.3 RQ3: Effect of Graph Encoding vs. Simple Projection

A natural question in structured knowledge integration is whether triplets should be treated as an unordered bag of facts or whether their relational structure should be explicitly modelled before injection into the expert. In the PT variant, outputs from the frozen triplet generator are mean-pooled into a single vector and projected into the expert's prefix space, preserving content but ignoring entity connections. In contrast, the PGT variant assembles accumulated triplets into a KG, applies GCN-based message passing to capture inter-triplet relations, and projects the resulting structure-aware embedding into the prefix (Sec. 3.4). As shown in Figure 4, PGT consistently outperforms PT across both iCRAFT-MD and iMEDQA. For example, Llama-3-8B improves from 55.5 % to 58.2 % in iCRAFT-MD and Llama-3-70B improves from 66.0 % to 71.3 % in iMEDQA. Gains are observed even for larger size LLMs such as Qwen-2.5-72B, suggesting

that the benefits are not simply due to limited capacity. These findings confirm that explicitly modelling inter-triplet relations through graph encoding enables the expert system to more effectively exploit structured clinical knowledge and reason across turns. Addressing RQ3, we find that explicitly modelling inter-triplet relations through graph encoding achieves higher diagnostic accuracy than simple projection, underlining the value of relational structure.

5.4 RQ4: Impact of Dialogue Turns on Expert Prediction Accuracy.

Understanding how performance scales with dialogue depth is crucial for evaluating whether a system can accumulate and reason over clinical evidence in interactive settings. To quantify the role of dialogue depth in our approach, we measure how diagnostic accuracy evolves as the consultation unfolds, up to a maximum of 10 dialogue turns. As shown in Figure 5, TriMediQ consistently outperforms MEDIQ in both iCRAFT-MD and iMEDQA, with the performance gap widening as dialogue length increases. For instance, on iCRAFT-MD, MEDIQ improves only modestly from 43.1% at turn 2 to 50.6% at turn 10. In contrast, TriMediQ rises from 52.2% to 58.2%. A similar trend is observed on iMEDQA, where the advantage of TriMediQ grows from early interactions and persists throughout later turns. In particular, MEDIQ saturates quickly (e.g., almost flat after turn 6), while TriMediQ continues to improve, suggesting that structured triplet chains enable more reliable long horizon reasoning by preserving and integrating evidence from earlier dialogue turns. With respect to RQ4, the evidence indicates that TriMediQ scales more effectively with dialogue depth than MEDIQ, maintaining improvements throughout longer consultations.

6 Conclusions

While Large Language Models excel in static and single-turn medical QA benchmarks, they struggle in interactive consultations where clinical facts are scattered across different dialogue sentences. To address this limitation, we proposed TriMediQ, a new approach that converts patient responses into triplets, aggregates them into a knowledge graph, and injects graph-encoded representations into a frozen LLM expert via a trainable projection. This design enables explicit multi-hop reasoning and enhances diagnostic accuracy across benchmarks and model scales. Our extensive experiments on two datasets, namely iCRAFT-MD and iMEDQA, using multiple LLM backbones, showed consistent and substantial improvements: TriMediQ outperforms recent MEDIQ by up to a 10.4% accuracy (Llama-3-70B in iMEDQA) and achieves robust gains even with smaller models such as +8.2% with Llama-3-8B in iCRAFT-MD. These results demonstrate that structured knowledge representations can be effectively aligned with frozen LLMs to improve accuracy in multi-turn clinical QA, offering a practical step toward enabling multi-hop reasoning in LLM-based medical assistants. Future work will extend this framework to richer ontologies, clinician-authored dialogues, and domains such as conversational decision support.

References

1. Almansoori, M., Kumar, K., Cholakkal, H.: Self-evolving multi-agent simulations for realistic clinical interactions. arXiv preprint arXiv:2503.22678 (2025)
2. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
3. Bornstein, B.H., Emler, A.C.: 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 (2001)
4. Cao, J., Fang, J., Meng, Z., Liang, S.: Knowledge graph embedding: A survey from the perspective of representation spaces. ACM Computing Surveys **56**(6), 1–42 (2024)
5. Church, K.W.: Word2vec. Natural Language Engineering **23**(1), 155–162 (2017)
6. Ding, H., Huang, B., Fang, Y., Liao, W., Jiang, X., Li, Z., Zhao, J., Wang, Y.: Promed: Shapley information gain guided reinforcement learning for proactive medical llms. arXiv preprint arXiv:2508.13514 (2025)
7. Fang, J., Meng, Z., Macdonald, C.: Kirag: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. Proceedings of the 63st Annual Meeting of the Association for Computational Linguistics (2025)
8. Han, T., Adams, L.C., Papaioannou, J.M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressem, K.K.: Medalpaca – an open-source collection of medical conversational ai models and training data (2023)
9. He, X., Tian, Y., Sun, Y., Chawla, N., Laurent, T., LeCun, Y., Bresson, X., Hooi, B.: G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. Advances in Neural Information Processing Systems **37**, 132876–132907 (2024)
10. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
11. Hoang, T.L., Sbodio, M.L., Galindo, M.M., Zayats, M., Fernandez-Diaz, R., Valls, V., Picco, G., Berrospi, C., Lopez, V.: Knowledge enhanced representation learning for drug discovery. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 10544–10552 (2024)
12. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
13. Jiang, B., Zhang, Z., Lin, D., Tang, J., Luo, B.: Semi-supervised learning with graph learning-convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11313–11320 (2019)
14. Jiang, Z., Zhong, L., Sun, M., Xu, J., Sun, R., Cai, H., Luo, S., Zhang, Z.: Efficient knowledge infusion via kg-llm alignment. Proceedings of the 62st Annual Meeting of the Association for Computational Linguistics (2024)
15. Jin, B., Liu, G., Han, C., Jiang, M., Ji, H., Han, J.: Large language models on graphs: A comprehensive survey. IEEE Transactions on Knowledge and Data Engineering (2024)
16. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (2021)

17. Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: PubMedQA: A dataset for biomedical research question answering. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2567–2577. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
18. Johri, S., Jeong, J., Tran, B.A., Schlessinger, D.I., Wongvibulsin, S., Cai, Z.R., Daneshjou, R., Rajpurkar, P.: Testing the limits of language models: A conversational framework for medical ai assessmentr. medRxiv (2023)
19. Johri, S., Jeong, J., Tran, B.A., Schlessinger, D.I., Wongvibulsin, S., Cai, Z.R., Daneshjou, R., Rajpurkar, P.: CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In: AAAI 2024 Spring Symposium on Clinical Foundation Models (2024)
20. Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., Park, H.W.: Mdagents: An adaptive collaboration of llms for medical decision-making. Advances in Neural Information Processing Systems **37**, 79410–79452 (2024)
21. Li, S.S., Balachandran, V., Feng, S., Ilgen, J.S., Pierson, E., Koh, P.W., Tsvetkov, Y.: Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
22. Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N.: Self-alignment pretraining for biomedical entity representations. 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2020)
23. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. arXiv preprint arXiv:2307.03172 (2023)
24. Masic, I.: Medical decision making-an overview. Acta Informatica Medica **30**(3), 230 (2022)
25. Meng, Z., Liu, F., Clark, T.H., Shareghi, E., Collier, N.: Mixture-of-partitions: Infusing large biomedical knowledge graphs into bert. arXiv preprint arXiv:2109.04810 (2021)
26. Meng, Z., Liu, S., Liang, S., Jani, B., Meng, Z.: Heterogeneous biomedical entity representation learning for gene–disease association prediction. Briefings in Bioinformatics **25**(5), bbae380 (2024)
27. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022)
28. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
29. Shang, T., He, W., Zheng, C., Li, L., Shen, L., Zhao, B.: Dynamicare: A dynamic multi-agent framework for interactive and open-ended medical decision-making. arXiv preprint arXiv:2507.02616 (2025)
30. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
31. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., y Arcas, B.A., Tomasev, N.,

Liu, Y., Wong, R., Semturs, C., Mahdavi, S.S., Barral, J., Webster, D., Corrado, G.S., Matias, Y., Azizi, S., Karthikesalingam, A., Natarajan, V.: Towards expert-level medical question answering with large language models (2023)
32. Su, X., Wang, Y., Gao, S., Liu, X., Giunchiglia, V., Clevert, D.A., Zitnik, M.: Kgarevion: an ai agent for knowledge-intensive biomedical qa. The Fourteenth International Conference on Learning Representations (2024)
33. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
34. Wang, H., Feng, S., He, T., Tan, Z., Han, X., Tsvetkov, Y.: Can language models solve graph problems in natural language? Advances in Neural Information Processing Systems **36**, 30840–30861 (2023)
35. Wang, Z., Liang, S., Liu, S., Meng, Z., Wang, J., Liang, S.: Sequence pre-training-based graph neural network for predicting lncrna-mirna associations. Briefings in Bioinformatics **24**(5), bbad317 (2023)
36. Xu, R., Cui, H., Yu, Y., Kan, X., Shi, W., Zhuang, Y., Jin, W., Ho, J., Yang, C.: 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 (2023)
37. Zhang, X., Wang, Y., Feng, Z., Chen, R., Zhou, Z., Zhang, Y., Xu, H., Wu, J., Liu, Z.: Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning. arXiv preprint arXiv:2506.12307 (2025)