MedKGI: Iterative Differential Diagnosis with Medical Knowledge Graphs and Information-Guided Inquiring

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have demonstrated significant promise in clinical diagnosis. However, current models struggle to emulate the iterative, diagnostic hypothesis-driven reasoning of real clinical scenarios. Specifically, current LLMs suffer from three critical limitations: (1) generating hallucinated medical content due to weak grounding in verified knowledge, (2) asking redundant or inefficient questions rather than discriminative ones that hinder diagnostic progress, and (3) losing coherence over multi-turn dialogues, leading to contradictory or inconsistent conclusions. To address these challenges, we propose MedKGI, a diagnostic framework grounded in clinical practices. MedKGI integrates a medical knowledge graph (KG) to constrain reasoning to validated medical ontologies, selects questions based on information gain to maximize diagnostic efficiency, and adopts an OSCE-format structured state to maintain consistent evidence tracking across turns. Experiments on clinical benchmarks show that MedKGI outperforms strong LLM baselines in both diagnostic accuracy and inquiry efficiency, improving dialogue efficiency by 30% on average while maintaining state-of-the-art accuracy.

1 Introduction

Large Language Models (LLMs) are increasingly demonstrating their value as powerful tools for clinical diagnosis (Wang et al., 2023; Singhal et al., 2025). However, real-world clinical reasoning is an iterative process in which doctors need to strategically construct diagnostic hypotheses and gather clinical information in a sequential manner to make the final decision. Current LLMs often struggle in such iterative, hypothesis-driven settings due to a fundamental discrepancy between their probabilistic, token-by-token generation and the systematic rigor required for clinical deduction (Hager et al., 2024). This gap leads to several critical limitations: a tendency to produce *hallucinations* by prioritizing plausible patterns over verified knowledge (Zhu et al., 2025); *ineffective questioning* due to the lack of an explicit reasoning framework (Li et al., 2025); and *context overloading* in multi-turn dialogues caused by their associative nature (Savage et al., 2024). As illustrated in Figure 1 (left), these limitations often result in redundant and non-strategic diagnostic dialogues by baseline LLMs, in contrast to the structured trajectory of rigorous medical reasoning.

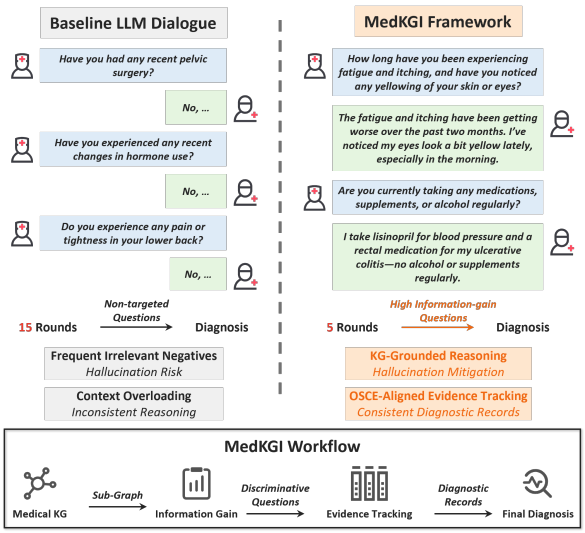

Figure 1: Comparison of diagnostic dialogues and the MedKGI workflow. Left: The dialogues from the baseline LLMs. Right: The dialogues from the proposed MedKGI framework. Bottom: The MedKGI workflow.

To bridge this gap, we ground our method in established differential diagnosis frameworks rather than language patterns. *Differential diagnosis* is inherently an iterative process of systematically weighing competing hypotheses against clinical evidence (Hannigen, 2018), which is the opposite of associative LLM reasoning. Accordingly, we integrate three key principles:

- **Knowledge-Anchored Hypothesis Generation:** Inspired by the clinical practice of grounding initial differentials in established medical knowledge (Zuo et al., 2025), we integrate a medical knowledge graph (KG) to generate diagnostic candidates based on verified disease–symptom relationships.
- **Strategic Uncertainty Reduction:** Following the differential diagnosis principle of prioritizing high-yield findings, we adopt an information gain-based questioning strategy (Liu et al., 2025) to select the most discriminative questions, thereby minimizing diagnostic uncertainty.
- **Iterative Evidence Refinement:** To simulate a doctor's belief updating process as new evidence emerges, we implement a state-tracking mechanism that maintains a coherent diagnostic record, enabling consistent hypothesis management while mitigating context overloading in long dialogues (Xu et al., 2024).

Building on these principles, we propose MedKGI, a diagnostic reasoning framework designed to emulate the systematic inquiry of human clinicians within the LLM paradigm. As shown in Figure 1, MedKGI integrates a medical knowledge graph to anchor all diagnostic reasoning in verified medical ontologies, thereby mitigating hallucinations. Building on this grounded knowledge (sub-graph), it employs an information gain–based question selection strategy. This strategy evaluates candidate questions by their expected reduction in diagnostic uncertainty, enabling MedKGI to prioritize the most discriminative inquiries and optimize diagnostic efficiency. Finally, MedKGI adopts the Objective Structured Clinical Examination (OSCE) format (Cushing et al., 2014) to maintain a structured diagnostic state. This state tracks and updates accumulated evidence across dialogue turns, which mitigates context overloading and ensures reasoning consistency.

Our key contributions are:

- **Information-Theoretic Formulation of Clinical Inquiry:** We formalize multi-turn diagnostic questioning as an information gain maximizing problem over a KG-constrained hypothesis subgraph, providing a principled criterion for question selection that existing LLMs lack.
- **Practical Approximations for Clinical Tractability:** We introduce *symptom specificity weighting* and *redundancy sharpening* to emphasize the diagnostic contribution of disease-specific symptoms while avoiding redundant questions, yielding an efficient approximation to the information gain objective.
- **Empirical Validation in MedKGI:** We instantiate the framework as MedKGI, coupling the information-gain criterion with a medical knowledge graph to ground reasoning in verified ontologies. Experiments show MedKGI outperforms state-of-the-art baselines while improving dialogue efficiency by 30%.

2 Related Work

We categorize existing clinical dialogue approaches into three paradigms.

LLM-Driven Sequential Diagnosis. Methods such as CoT prompting (Dai et al., 2025), domain-specialized models Huatuo (Wang et al., 2023) and Meditron (Chen et al., 2023), and multi-agent collaboration MedAgent (Kim et al., 2025b) and PATIENCE (Zhu et al., 2025) leverage LLM reasoning for diagnosis, but remain prone to hallucinations in multi-turn scenarios due to absent external grounding (Zuo et al., 2025).

Knowledge-Augmented Approaches. RAG-based methods (MRD-RAG, ClinicalRAG) (Sun et al., 2025; Lu et al., 2024) and KG-grounded reasoning such as Unit of Thought (UoT) (Hu et al., 2024) integrate external knowledge, yet treat retrieval statically without dynamic state-tracking across turns (Wang et al., 2025).

Agent-based Clinical Frameworks. Multi-agent systems such as DDO (Jia et al., 2025b), MeDDxAgent (Rose et al., 2025), MEDIQ (Li et al., 2024), CoD (Chen et al., 2025a) and DoctorAgent-RL (Feng et al., 2025) decompose clinical workflows across specialized agents, but lack principled question selection criteria, relying on heuristic role-playing rather than information-theoretic objectives to optimally reduce diagnostic uncertainty (Chen et al., 2025b).

Methodological Differentiation. Unlike prior works that augment LLM prompts with static KG retrieval, MedKGI integrates three mechanisms for multi-turn consultation:

- Dynamic subgraph construction bounded by the LLM-proposed candidate diseases at each turn.
- Iterative Bayesian Belief Update through KG-verified symptom-disease priors.

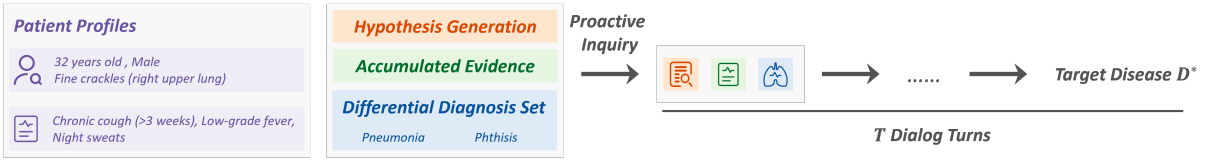

Figure 2: An illustration of the iterative hypothesis refinement process and its corresponding clinical differential diagnosis example.

- Information gain-based inquiry over the KG to maximize uncertainty reduction.

3 Problem Definition

The multi-step clinical diagnosis can be modeled as an iterative decision-making process that refines hypotheses over T dialogue turns (Figure 2). The process begins with the patient profile $\mathcal{P}$, which includes demographics and chief complaints. Over a sequence of dialogue turns t, the framework maintains a dynamic clinical state. At turn t, given $\mathcal{P}$ and accumulated evidence $\mathcal{E}_t$, the objective is to estimate the posterior probability for each candidate disease $D \in \mathcal{D}_t$, where $\mathcal{D}_t = \{D_1, D_2, \ldots, D_n\}$ through evidence collection. The proactive symptom inquiry is defined as identifying the optimal inquiry s that maximizes the expected reduction in diagnostic uncertainty. This mechanism enables the framework to iteratively generate and refine D_t until the target disease D^* is reached:

$$P(D \mid \mathcal{P}, \mathcal{E}_t) \rightarrow \delta(D, D^*)$$

4 Methodology

As illustrated in the differential diagnosis scenario (Figure 2), given patient profiles and symptoms, the objective is to narrow a differential diagnosis set toward the target disease D^* through sequential evidence collection. Unlike static classification, this scenario requires actively navigating a hypothesis space, simulating a doctor's cognitive process (Polotskaya et al., 2024). Specifically, we formulate the iterative refinement process (Figure 3) as a knowledge-guided active diagnostic framework, where a doctor agent iteratively refines the differential diagnosis set by strategically gathering discriminative evidence.

To realize this iterative and knowledge-driven process, we design the MedKGI workflow integrating three components: (1) **Entity Extraction & Alignment:** maps the diagnosis input and generated hypothesis to a medical KG, constructing a diagnostic subgraph grounded in clinically validated disease-symptom relationships to mitigate hallucinations, (2) **Information Gain-Based Inquiry:** calculates information gain to identify discriminative symptoms, ensuring each clinical inquiry maximally reduces diagnostic uncertainty and improves diagnostic efficiency, (3) **OSCE-Aligned Diagnostic Record Management:** organizes accumulated evidence into an OSCE-format diagnostic record, maintaining a consistent state to prevent context overloading.

4.1 Diagnosis Workflow of MedKGI

At any diagnostic turn t, the doctor agent receives the following as input:

- The *Patient Profile* (e.g., age, sex, and chief complaints) provided at the beginning;
- The *Patient's Recent Utterance*, which may contain new symptoms or responses to prior questions;
- The *Accumulated Diagnostic Record*, a structured OSCE-aligned summary (detailed in Section 4.4) containing confirmed symptoms, medical history, and examination findings up to turn $t-1$.

Based on this context, the doctor agent generates two key outputs:

- A natural language *Clinical Inquiry* to elicit discriminative evidence;
- A *Differential Diagnosis Set* outputted by doctor agent;

To ensure efficiency and diagnostic precision, diagnostic process concludes when one of the following termination conditions is met

- **Turn Limit.** If the dialogue reaches T_{max} turns, the doctor agent must issue a final diagnosis.
- **Stagnation Detection.** If the differential diagnosis set remains unchanged for n consecutive turns, doctor agent is prompted to seek evidence that could refute the current hypothesis. If no such symptom exists, the doctor agent outputs the final diagnosis.

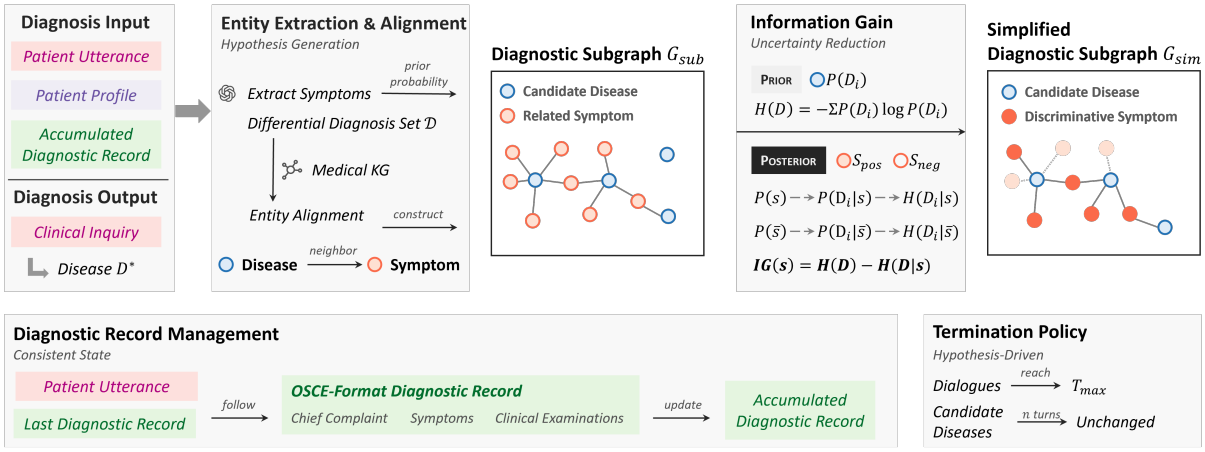

Figure 3: An overview of MedKGI framework. Given a patient's chief complaint, MedKGI iteratively refines differential diagnosis through (1) medical knowledge graph alignment, (2) information gain–driven symptom inquiry to minimize diagnosis uncertainty, and (3) OSCE-aligned diagnostic records for coherent evidence tracking. A hypothesis-driven termination policy ensures diagnostic efficiency.

4.2 Entity Extraction and Knowledge Graph Alignment

At each turn, the LLM first proposes a preliminary differential diagnosis set, which is then mapped to the medical KG. To align medical terms mentioned in the patient utterance with standardized entities in the KG, we implement a multi-stage alignment pipeline:

- **Exact Matching.** For standard medical terms in the dialogue, we query the KG for an entity that exactly matches the candidate disease name.
- **Edit-Distance Matching.** For minor spelling variations or errors, we apply the Levenshtein Distance (Levenshtein, 1965) to identify approximate matches, allowing a maximum edit distance of 3.
- **Semantic Embedding Matching.** For conceptually equivalent but lexically divergent expressions, we leverage the pre-trained PubMedBERT (Gu et al., 2021) to generate vector embeddings for the candidate disease name and all KG disease entity names. We calculate cosine similarity and retrieve the top-ranked entity, discarding matches with a similarity score below a threshold $\tau = 0.85$ to ensure alignment quality.

4.3 Information Gain-Based Symptom Selection

Once the candidate diseases are mapped to the KG, we construct a task-specific diagnostic subgraph $G_{sub} = (V_{sub}, E_{sub})$, where $V_{sub} = \bigcup_{i=1}^{n}\{D_i\} \cup N(D_i)$ consists of differential diagnosis set $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$ and all symptoms connected to them in the KG: $N(D_i) = \{s \in V_{KG} \vee s \in \mathcal{S} : (D_i, s) \in E_{KG}\}$. We intentionally restrict $N(D_i)$ to 1-hop neighbors for methodological and practical reasons. Medical KGs already encode clinically relevant manifestations as direct disease–symptom edges, so 1-hop neighborhoods are typically sufficient for differential diagnosis. Extending retrieval to multiple hops, in contrast, tends to introduce non-discriminative symptoms and dilute the Information Gain signal. A 2-hop experiment on PrimeKG further confirmed the practical cost, where the retrieved subgraph expanded to over 5,000 nodes and 10,000 edges per query, with retrieval latency exceeding 400 seconds, which is far beyond the response budget required for interactive diagnosis. To strategically reduce diagnostic uncertainty, MedKGI selects symptom queries that maximize information gain (Quinlan, 1986). This selection is made over the diagnostic subgraph G_{sub} and is conditioned on the patient's reported positive and negative symptoms (S_{pos}, S_{neg}).

4.3.1 Posterior Disease Probability Estimation

First, we establish the context by constructing a diagnostic subgraph G_{sub}, where $V_{sub} = \bigcup_{i=1}^{n}\{D_i\} \cup N(D_i)$ consists of the differential diagnosis set $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$ and all symptoms connected to them in the KG: $N(D_i) = \{s \in V_{KG} \vee s \in \mathcal{S} : (D_i, s) \in E_{KG}\}$. We initialize the prior probability of each candidate disease D_i based on its average semantic similarity to the confirmed symptoms $S_{pos} \subseteq \mathcal{S}$ extracted from the

current dialogue:

$$P(D_i) = \frac{1}{\mid S_{pos} \mid} \sum_{S \in S_{pos}} semantic_sim(s, D_i)$$

where $semantic_sim(\cdot, \cdot)$ denotes cosine similarity between the PubMedBERT (Gu et al., 2021) embeddings of symptom $s \in S_{pos}$ and disease D_i.

Given the accumulated observed symptoms S_{pos} and S_{neg}, we update disease beliefs over candidate disease $\mathcal{D}$ using Bayes' theorem:

$$P(D_i \mid S_{pos}, S_{neg}) = \frac{P(S_{pos}, S_{neg} \mid D_i) \cdot P(D_i)}{P(S_{pos}, S_{neg})}$$

Assuming conditional independence among symptoms given a disease, the likelihood factorizes as:

$$P(S_{pos}, S_{neg}) = \sum_{j=1}^{n} P(S_{pos}, S_{neg} \mid D_j) \cdot P(D_j)$$

where we adopt a uniform conditional probability model: $P(s \mid D_i) = 1/\,|N(D_i)|$ for all symptoms $s \in N(D_i)$.

4.3.2 Information Gain Computation

For the current differential disease set $\mathcal{D}$ with posterior probabilities $P(D_i \mid S_{pos}, S_{neg})$, we compute the prior diagnostic uncertainty using the Shannon Entropy:

$$H(\mathcal{D}) = -\sum_{i=1}^{n} P(D_i) \log P(D_i)$$

For any symptom s, we compute its marginal probability and the resulting posterior disease distributions:

$$P(s) = \sum_{i=1}^{n} P(s \mid D_i) P(D_i),$$

$$P(D_i \mid s) = \frac{P(s \mid D_i) P(D_i)}{P(s)}$$

Finally, the Information Gain of asking about symptom s is defined as the expected reduction in entropy:

$$IG(s) = H(\mathcal{D}) - H(\mathcal{D} \mid s)$$

where $H(\mathcal{D} \mid s)$ represents the expected conditional entropy after observing symptom s:

$$H(\mathcal{D} \mid s) = P(s) H(\mathcal{D} \mid s^+) + P(\neg s) H(\mathcal{D} \mid s^-)$$

and $H(\mathcal{D} \mid s^+)$ and $H(\mathcal{D} \mid s^-)$ are the entropies if the symptom is observed positive or negative, respectively. The doctor agent then selects the top-k symptoms with the highest $IG(s)$ for strategic inquiry, ensuring that each subsequent question maximally reduces uncertainty. Compared to methods that rely on predefined question templates or fixed inquiry sequences, MedKGI dynamically adapts questions based on the evolving diagnostic hypothesis, enabling more targeted and efficient information gathering.

4.3.3 Symptom Specificity and Redundancy Mitigation

Our base formulation based on Naive Bayes introduces two biases: (1) uniform treatment of symptoms regardless of clinical specificity, causing the agent to frequently select low-specificity questions and waste inquiey turns, and (2) redundant information gain from correlated symptoms rendering the posterior overly diffuse and slow convergence to the correct diagnosis We introduce two mechanisms to address these biases without requiring joint distribution modeling.

Symptom Specificity Weighting. Analogous to IDF in information retrieval, we define an Inverse Disease Frequency weight $w(s)$ for a symptom s:

$$w(s) = \log \frac{|\mathcal{D}|}{|\mathcal{N}(s)|}$$

where $|\mathcal{D}|$ is the total number of diseases in the knowledge graph, and $|\mathcal{N}(s)|$ is the number of diseases associated with s, ensuring that MedKGI prioritizes symptoms with higher diagnostic specificity among those with similar IG.

Redundancy Sharpening. To discourage redundantly inquiring about symptoms similar to already-explored ones, we define

$$\gamma(s) = 1 - \max_{s' \in S_{pos}} \mathrm{sim}(s, s')$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes the cosine similarity between symptom embeddings, bounded in $[0, 1]$.

Combined Formulation. Integrating both mechanisms, the adjusted marginal symptom probability becomes:

$$P(s) = \sum_{i=1}^{n} w(s) \cdot P(s \mid D_i)^{\gamma(s)} P(D_i)$$

where $w(s)$ penalizes uninformative common symptoms to address the uniform probability

bias, while $\gamma(s)$ dynamically suppresses correlated symptoms to address the independence assumption bias.

4.4 Consistent State by Diagnostic Record Management

To support coherent and consistent reasoning, we employ the LLM to generate and maintain a structured diagnostic record in JSON format, aligned with the Objective Structured Clinical Examination (OSCE) standard. At the beginning of each dialogue session, we initialize an empty diagnostic record following a predefined schema, including chief complaint, symptoms, and recent medical examinations.

At each turn, MedKGI takes the latest diagnostic record and patient utterance as input. Then, MedKGI outputs an updated diagnostic record that integrates new evidence while preserving clinical context. This accumulated diagnostic record prevents context overloading across turns, which commonly occurs when context windows accumulate redundant information in vanilla prompting methods (Schmidgall et al., 2024).

5 Evaluation

5.1 Experiment Setup

Datasets. We conducted experiments on two medical QA benchmarks: MedQA (Jin et al., 2021) and CMB (Wang et al., 2024; cme, 2023). While real-world consulation datasets (e.g., MedDG(Liu et al., 2022)) provide naturalistic dialogue, they typically lack expert-validated diagnostic ground truth, making quantitative evaluation of diagnostic accuracy infeasible. Following the settings of AgentClinic (Schmidgall et al., 2024), we simulated the multi-agent medical consultation scenarios based on the cases from MedQA and CMB. To prevent the evaluation from degenerating into scripted interactions where the Doctor Agent passively receives pre-provided hints in exam questions, we implemented an information-hiding mechanism. Specifically, the description from the original multiple-choice questions is converted to the Patient Agent's profile. The Doctor Agent initially receives only a chief complaint and must strategically formulate inquiries to uncover hidden diagnostic clues step-by-step. We denote the processed datasets as Agent-MedQA and Agent-CMB. To further assess multi-modal clinical reasoning, we additionally introduce a dataset of real-world cases from the NEJM Image Challenge[1].

Baselines. We compared MedKGI against 12 baselines across four categories: (1) Dialog-Based Methods, (2) KG-Based Methods, (3) Agent-Based Methods, and (4) SFT-Based Methods. Specialized medical LLMs and SFT variants are excluded from NEJM evaluation due to lacking multimodal capabilities or training data. Full baseline descriptions are provided in Appendix A.1.

Implementation. To simulate realistic clinical interactions, we implemented a multi-agent framework with three specialized agents, adapted from AgentClinic (Schmidgall et al., 2024):

- Doctor agent asks up to $T_{max} = 20$ questions.
- Patient agent responds only with symptom descriptions and never reveals diagnosis.
- Measurement agent simulates the outcome of laboratory tests or medical examinations.

To ensure fair comparison, all baselines are evaluated under fully matched experimental conditions, interacting with the same simulated agent and is assessed using the same criteria.

We employed Qwen3-8B (Yang et al., 2025) and Llama3.1-8B-Instruct (AI@Meta, 2024) for Agent-MedQA and AgentCMB, and Qwen3-VL-8B-Instruct (Yang et al., 2025) for NEJM. Details on KG integration, prompts and configurations are provided in the Appendix B.4.

Metrics. Diagnostic accuracy (*acc*): exact match between the final diagnostic output and the ground truth D^*. Dialogue rounds (*Rounds*): average turns to reach a diagnosis for each method. Fewer *Rounds* indicate more efficient diagnosis.

5.2 Main Result

Table 1 presents a comprehensive comparison of MedKGI against state-of-the-art baselines across three medical consultation benchmarks, AgentMedQA, AgentCMB, and NEJM, using multiple backbone LLMs. Our method demonstrates superior performance in both diagnostic accuracy and efficiency.

Overall Performance. MedKGI achieves superior accuracy across three benchmarks using comparable base models: 69.81% on AgentMedQA (Qwen3-8B), 68.21% on AgentCMB (Qwen3-8B), and 69.09% on NEJM (Qwen3-VL-8B-Instruct). Notably, these results are obtained with the fewest dialogue rounds, 9.11, 9.13, and 10.53 out of a maximum of 20 rounds respectively.

[1] https://www.nejm.org/image-challenge

Table 1: Comprehensive evaluation of MedKGI across three benchmarks Agent-MedQA, AgentCMB, and NEJM using Qwen3-8B, Llama3.1-8B-Instruct, Qwen3-VL-8B-Instruct as base language models. All methods are evaluated with a maximum dialogue round of 20. Reported metrics including average dialogue rounds (*Rounds* ↓) and diagnostic accuracy (*Acc* (%) ↑). Methods marked with * employ specialized LLMs (i.e.HuatuoGPT-o1, Meditron-7B, and DiagnosisGPT-7B) rather than the base LLM used in our unified evaluation. Best results per column are **bolded**; second-best are underlined.

Baselines	**Agent-MedQA**				**Agent-CMB**				**NEJM**	
	Qwen3-8B		Llama3.1-8B-Instruct		Qwen3-8B		Llama3.1-8B-Instuct		Qwen3-VL-8B-Instruct	
	Rounds	*Acc (%)*	*Rounds*	*Acc (%)*	*Rounds*	*Acc (%)*	*Rounds*	*Acc (%)*	*Rounds*	*Acc (%)*
LLM-Based										
AgentClinic	11.32	59.43	10.37	50.00	10.00	58.28	11.23	59.60	13.28	54.55
CoT	18.92	24.52	17.46	34.90	18.32	43.05	16.33	50.33	16.46	50.91
Huatuo* *(HuatuoGPT-o1)*	16.70	56.60	-	-	16.56	62.25	-	-	-	-
Medical-CoT* *(MediTron-7B)*	18.94	60.37	-	-	18.34	66.23	-	-	-	-
KG-Based										
MCTS-BT	14.20	45.28	14.98	39.62	13.78	54.97	14.21	42.38	11.50	56.36
MCTS-MV	14.63	53.77	12.86	52.83	14.77	60.26	12.91	56.95	13.95	67.27
UoT	11.47	54.71	11.01	51.89	10.71	64.28	10.96	58.94	12.32	54.55
Agent-Based										
MediQ	13.80	61.32	13.85	49.06	13.76	65.56	15.01	**60.93**	14.48	65.45
DDO	17.27	61.32	18.39	50.94	17.38	63.58	18.01	**60.93**	17.91	**70.73**
MEDDxAgent	16.44	60.38	16.02	49.06	16.09	61.59	16.48	57.62	16.36	65.45
CoD* *(DiagnosisGPT-7B)*	13.32	56.60	-	-	11.99	28.50	-	-	-	-
SFT-Based										
SFT	11.51	51.89	11.21	49.06	12.04	59.60	9.95	52.98	-	-
SFT-GT	9.35	50.94	10.27	40.57	9.93	55.63	10.26	51.66	-	-
Ours	**9.11**	**69.81**	**10.20**	**53.77**	**9.13**	**68.21**	**9.72**	60.26	**10.53**	69.09

Comparison with LLM-based Methods. Compared to general LLM-based approaches, MedKGI outperforms methods like AgentClinic and CoT across all benchmarks. It also surpasses specialized medical LLMs (marked with *). For instance, on AgentCMB, MedKGI using Qwen3-8B achieves higher accuracy (68.21%) than Medical-CoT with MediTron-7B (66.23%), while doing so in significantly fewer dialogue rounds.

Comparison with KG-based and Agent-based Methods. Among KG-based methods, MedKGI surpasses even competitive approaches like MCTS-MV and UoT. For instance, on AgentCMB with Qwen3-8B, it achieves 68.21% accuracy, exceeding UoT's 64.28%. In contrast to agent-based approaches such as MediQ, DDO, and MEDDxAgent, MedKGI also demonstrates superior performance. Furthermore, compared to the state-of-the-art method, MedKGI achieves comparable or better accuracy while typically requiring far fewer rounds across all three datasets.

Comparison with SFT-based Methods. While SFT-based methods achieve competitive dialogue efficiency, their accuracy lags behind that of our method. For example, on AgentMedQA using Qwen3-8B, SFT-GT achieves comparable efficiency (9.35 average rounds vs. our 9.11) but its accuracy (50.94%) is significantly lower than ours (69.81%).

5.3 Analysis of Superior Performance

The performance of MedKGI generalizes across different backbone LLMs. For example, with Llama3.1-8B-Instruct, our method achieves the

highest accuracy on AgentMedQA (53.77%) and the second-highest on AgentCMB (60.26%), while consistently requiring the fewest dialogue rounds. The superiority of MedKGI across diverse benchmarks and LLMs stems from three key factors: knowledge grounding, context-aware reasoning, and efficient inquiry. First, unlike methods that rely solely on pre-trained LLM knowledge (e.g., AgentClinic, CoT), which may lack structured clinical reasoning, MedKGI integrates a medical knowledge graph (KG). This external grounding enables precise inference and provides a structured hypothesis space for active querying (Jia et al., 2025a). Our ablation study (Table 3) confirms that removing the KG leads to a significant drop in accuracy (-25.47% on AgentMedQA). Second, compared to other KG-based methods that often rely on heuristic metrics for symptom selection, MedKGI selects questions based on information gain that accounts for patient-specific context. This approach avoids both random noise and popularity bias, leading to more discriminative queries (Kim et al., 2025a). Third, in contrast to agent-based methods (e.g., DDO, MEDDxAgent), our framework minimizes redundant interactions by dynamically pruning the candidate symptom set based on information gain and maintaining diagnostic records. This enables MedKGI to achieve diagnosis in fewer rounds while maintaining high symptom coverage.

5.4 Error Case Analysis

We conducted an error analysis on 60 failure cases, categorizing them into five types: (1) critical information omission, (2) premature conclusion, (3) inefficient inquiry patterns, (4) evidence misinterpretation, and (5) knowledge gaps. Errors were predominantly concentrated in inefficient inquiry patterns (55/60) and evidence misinterpretation (24/60). To address the dominant failure mode of inefficient inquiry, we introduced Symptom Specificity Weighting and Redundancy Sharpening to prioritize high-discriminability symptoms and suppress redundant questions, respectively.

5.5 Effect of Specificity Weighting and Redundancy Sharpening

To empirically validate the contribution of the Symptom Specificity Weigh $w(s)$ introduced in Section 4.3.2, we conducted a comparative analysis between the base information gain-based strategy and specificity-weighted strategy. The evaluation focuses on two key dimensions: diagnostic efficiency and diagnostic accuracy.

Improvement of Diagnostic Accuracy and Efficiency. The symptom specificity weighting effectively enhances both diagnostic accuracy and efficiency. In medical scenarios where multiple diseases share common symptoms, relying on generic inquiries often leads to diagnostic drift(Gigerenzer and Gaissmaier, 2011), where the framework trapped in non-discriminative data collection process. In contrast, our Weighted IG strategy prioritizes discriminative symptoms in the early stages of diagnosis, accelerating convergence toward the correct diagnosis. The diagnostic accuracy improves from 69.81% to 70.76% on the Agent-MedQA Dataset and from 68.21% to 69.54% on the Agent-CMB Dataset. We also observed Weighted IG reduces dialogue rounds by prioritizing pathognomonic symptoms over ubiquitous ones, rapidly narrowing the differential diagnosis.

Qualitative Observation. We conducted a case study where the ground truth was epidermoid cyst. Our framework without symptom specificity weighting exhibited redundancy by repeatedly inquiring about generic symptoms (e.g., pain and swelling). Instead, it prematurely pivoted to lateral neck pathologies, ultimately requiring 20 dialogue turns to reach a conclusion with an incorrect diagnosis. In contrast, the method incorporating specificity weighting demonstrated superior clinical discriminative awareness. It inquiry key diagnostic features, including texture, skin appearance, and discharge, which align with the classic presentation of epidermoid cyst, confirming that our TF-IDF-inspired formulation effectively injects clinical sensitivity into the diagnostic dialogue process.

6 Conclusion

In this work, we present MedKGI, a framework that formalize multi-step clinical diagnosis as an active, knowledge-guided refinement process, combining KG-based hypothesis grounding, information gain–driven inquiry strategy and a structured diagnostic record aligned with clinical standards. Unlike existing approaches that rely on static retrieval, heuristic questioning, or ungrounded LLM reasoning, our framework explicitly models the evolving clinical state and optimizes each diagnostic step toward maximal discriminative power. Experimental results demonstrate that MedKGI achieves both superior diagnostic accuracy and dialogue efficiency.

Limitations

While MedKGI demonstrates promising performance in differential diagnosis, there are several limitations that warrant discussion. First, our patient simulation relies on LLM-generated case descriptions that may not fully capture the ambiguity of real patient narratives. Critically, our patient agent assumes cooperative and coherent symptom reporting, reflecting an idealized clinical interaction. In reality, patients often exhibit cognitive or linguistic biases: underreporting stigmatized symptoms, inaccurate recall, or anxiety-driven concerns rather than physiological reasoning. In addition, our information gain computations assume conditional independence among symptoms given a disease and employ uniform likelihood over symptom–disease edges in KG. This assumption may lead to suboptimal question selection when diseases are distinguished primarily by complex symptom patterns.

Ethical Consideration

The application of AI in diagnosis support has raised ethical concerns that we carefully acknowledge. MedKGI is proposed as a diagnostic reasoning assistant rather than a replacement for licensed doctors. All outputs must be validated by human doctors before any clinical action is taken. Our evaluation data are derived from publicly available and anonymized datasets AgentMedQA, AgentCMB and NEJM Image Challenge. No real patient records are used, ensuring compliance with privacy standards.

References

2023. Cmb: Chinese medical benchmark. https://github.com/FreedomIntelligence/CMB.

AI@Meta. 2024. Llama 3 model card.

LLM Benchmarks. 2026. gpt4 by openai benchmarks. https://llm-benchmarks.com/models/openai/gpt4.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2025a. COD, Towards an Interpretable Medical Agent Using Chain of Diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14345–14368.

Xi Chen, Huahui Yi, Mingke You, Weizhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, and 1 others. 2025b. Enhancing Diagnostic Capability with Multi-Agents Conversational Large Language Models. *npj Digital Medicine*, 8(1):159.

Zeming Chen, Alejandro Hern A Ndez Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70B: Scaling Medical Pretraining for Large Language Models. *arXiv e-prints*, page arXiv:2311.16079.

A. M. Cushing, J. S. Ker, P. Kinnersley, P. Mckeown, J. Silverman, J. Patterson, and O. M. R. Westwood. 2014. Objective Structured Clinical Examination. APA PsycTests Database Record.

Guangxin Dai, Xiang Li, Lizhou Fan, and Xin Ma. 2025. Enhancing Medical Diagnostic Reasoning with Chain-of-Thought in Large Language Models. In *2025 International Conference on Mechatronics, Robotics, and Artificial Intelligence*, pages 294–299.

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 e-prints*, page arXiv:2505.19630.

Gerd Gigerenzer and Wolfgangyr Gaissmaier. 2011. Heuristic Decision Making. *Annual Review of Psychology,*, 62:451.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1).

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024. Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-making. *Nature Medicine*, 30(9):2613–2622.

Sarah Hannigen. 2018. Differential Diagnosis. In *Encyclopedia of Clinical Neuropsychology*, pages 1148–1148.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of Thoughts: Uncertainty-Aware Planning Enhances Information Seeking in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 24181–24215.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025a. medIKAL: Integrating Knowledge Graphs as Assistants of LLMs for Enhanced Clinical Diagnosis on EMRs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9278–9298.

Zhihao Jia, Mingyi Jia, Junwen Duan, and Jianxin Wang. 2025b. DDO: Dual-Decision Optimization for LLM-Based Medical Consultation Via Multi-Agent Collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26380–26397.

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).

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025a. Limitations of Large Language Models in Clinical Problem-Solving Arising From Inflexible Reasoning. *Scientific Reports*, 15(1):39426.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel Mcduff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2025b. MDAgents: An Adaptive Collaboration of Llms for Medical Decision-Making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452.

Vladimir I. Levenshtein. 1965. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet physics. Doklady*, 10:707–710.

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 *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Pedram Hosseini, Bryceton G. Thomas, Jessica M. Sin, Bing Ren, Jonathan S. Ilgen, Yulia Tsvetkov, and Maarten Sap. 2025. Alfa: Aligning LLMs to Ask Good Questions a Case Study in Clinical Reasoning. In *Second Conference on Language Modeling*.

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Yining Hua, Peilin Zhou, and 1 others. 2025. Application of Large Language Models in Medicine. *Nature Reviews Bioengineering*, 3(6):445–464.

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 *Natural Language Processing and Chinese Computing*, pages 447–459.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024. ClinicalRAG: Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 64–68.

Kristina Polotskaya, Carlos S. Muñoz-Valencia, Alejandro Rabasa, Jose A. Quesada-Rico, Domingo Orozco-Beltrán, and Xavier Barber. 2024. Bayesian Networks for the Diagnosis and Prognosis of Diseases: A Scoping Review. *Machine Learning and Knowledge Extraction*, 6(2):1243–1262.

J. R. Quinlan. 1986. Induction of Decision Trees. *Machine Learning*, 1(1):81–106.

Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. 2025. MEDDxAgent: A Unified Modular Agent Framework for Explainable Automatic Differential Diagnosis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13803–13826.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine*, 7(1):20.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: A Multimodal Agent Benchmark to Evaluate Ai in Simulated Clinical Environments. *arXiv e-prints*, page arXiv:2405.07960.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward Expert-Level Medical Question Answering with Large Language Models. *Nature Medicine*, 31(3):943–950.

Penglei Sun, Yixiang Chen, Xiang Li, and Xiaowen Chu. 2025. The Multi-Round Diagnostic RAG Framework for Emulating Clinical Reasoning. *arXiv e-prints*, page arXiv:2504.07724.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning Llama Model with Chinese Medical Knowledge. *arXiv e-prints*, page arXiv:2304.06975.

Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2025. Adaptive Retrieval-Augmented Generation for Conversational Systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 491–503.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, and 1 others. 2024. CMB: A Comprehensive Medical Benchmark in Chinese. In *Proceedings of the 2024 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6184–6205.

Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024. Reasoning Like a Doctor: Improving Medical Dialogue Systems via Diagnostic Reasoning Process Alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6796–6814.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 Technical Report. *arXiv e-prints*, page arXiv:2505.09388.

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. *arXiv e-prints*, page arXiv:2502.07143.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. Can We Trust AI Doctors? A Survey of Medical Hallucination in Large Language and Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2025. KG4Diagnosis: A Hierarchical Multi-Agent LLM Framework with Knowledge Graph Enhancement for Medical Diagnosis. In *Proceedings of The First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281, pages 195–204.

A Implementation Details

A.1 Base Model and Inference Configuration

- Base models: Qwen3-8B, Meta-Llama-3.1-8B-Instruct, and Qwen3-VL-8B-Instruct
- Inference temperature: 0.05
- Max tokens: 2048

A.2 KG Implementation

To provide a clinically grounded foundation for our framework, we utilize PrimeKG (Precision Medicine Knowledge Graph) (Chandak et al., 2023), which is a comprehensive resource that integrates over 20 high-quality primary sources, including Orphanet, Mayo Clinic, and DrugBank.

PrimeKG comprises:

- Nodes: Approximately 17,000 disease entities and 1,300 symptom entities.
- Edges: We prioritize two primary relationship types:
  - Disease-Symptom: Indicating clinical manifestations associated with specific pathologies.
  - Disease-Disease: Representing comorbidity links and hierarchical relationships (e.g., "is-a" or "part-of" relations) that assist in differential grouping.

A.3 LoRA Fine-Tuning Hyperparameters

We fine-tune the base LLM using Low-Rank Adaptation (LoRA) (Hu et al., 2022) with the following configuration:

- Learning rate: 2e-4
- Batch size:
  - `per_device_train_batch_size = 2`
  - `gradient_accumulation_steps = 4`
  - `Effective batch size = 2 × 4 = 8`
- LoRA rank (`r`): 8
- Target modules: `["q_proj", "k_proj", "v_proj", "o_proj"]`
- LoRA alpha: 32
- Dropout: 0.1
- Training epochs: 3
- Warmup steps: 1,000

B Evaluation Details

B.1 Case Study of KG-Constrained Diagnostic Reasoning

To illustrate how MedKGI leverages the medical KG to ground, we present a representative case drawn from Agent-MedQA.

Initial Complaint. The patient first reports:

> *"He's been turning away from the bottle, and when he does take it, he only drinks a little before getting fussy and arching his back. He hasn't vomited or choked, but he seems uncomfortable during feeds—like he's in pain."*

KG-Constrained Candidate Generation. MedKGI first aligns the inferred candidate diseases to KG nodes, including Hirschsprung disease, gastroesophageal reflux disease, cystic fibrosis, congenital hypothyroidism, and infantile hypertrophic pyloric stenosis. It then constructs a diagnostic subgraph around these diseases, retrieving associated symptoms such as Gastroesophageal reflux, Abdominal pain, Nausea and vomiting, Failure to thrive in infancy, Weight loss, Intestinal obstruction, Malabsorption, and Feeding difficulties.

KG-Guided Follow-up Inquiry. Based on this diagnostic subgraph, MedKGI identify symptoms with the highest discriminative value, such as Abdominal pain (0.984), Failure to thrive in infancy (0.950), Nausea and vomiting (0.949). Guided by these KG-supported clues, the Doctor Agent issues the following targeted inquiry:

> *"How long has this feeding refusal and distress been going on, and has it been getting progressively worse?"*

This question is designed to separate a benign presentation from a gastrointestinal disorder. A worsening trajectory accompanied by abdominal discomfort, poor growth, constipation, or obstruction-related features would constitute substantially stronger evidence for Hirschsprung disease than isolated reflux symptoms.

Takeaway. This case highlights two key features of MedKGI: (i) by anchoring candidate generation to the KG, MedKGIavoids being misled by superficial symptom similarity and keeps the correct diagnosis, (ii) by using KG structure to select maximally informative follow-up questions, it guidess the consultation toward evidence that discriminates among candidate diseases, rather than merely confirming the most obvious one. Together, these features illustrate how KG-constrained reasoning assists efficient diagnosis.

B.2 Detailed Description of Baselines

We provide a comprehensive overview of the 12 baseline methods used in our experiments:

- **Dialog-Based Methods.** **AgentClinic** (Schmidgall et al., 2024): Simulates clinician–patient dialogues for diagnosis. CoT (Chain-of-Thought): Appends "Let's think step by step" to encourage explicit reasoning. Huatuo (Wang et al., 2023) and (Chen et al., 2023): Representative specialized medical LLMs.
- **KG-Based Methods.** MCTS-BT: Uses Monte Carlo Tree Search with backtracking for hypothesis refinement. MCTS-MV: Extends MCTS by ranking symptom queries based on contextual informativeness. UoT (Unit of Thought) (Hu et al., 2024): Constructs symptom-centric "units" around confirmed positive symptoms and prioritizes structural importance.
- **Agent-Based Methods.** MEDIQ (Li et al., 2024): A diagnostic agent implementing multiple diagnostic strategies through sequential dialogues. CoD (Chen et al., 2025a): Selects questions by maximizing information entropy over candidate diseases. DDO (Jia et al., 2025b): a multi-agent framework that dynamically chooses symptoms using diverse strategies. MEDDxAgent (Rose et al., 2025): Adapts questioning strategy based on diagnostic uncertainty.
- **SFT-Based Methods.** SFT / SFT-GT: Fine-tuning on Qwen3-8B using generated dialogues by AgentClinic (Schmidgall et al., 2024), with or without ground-truth labels respectively.

B.3 Supplementary Experiment on MedDG

To further validate the generalizability of our framework, we conducted a supplementary experiment on MedDG (Liu et al., 2022), a large-scale Chinese medical dialogue dataset collected from online consultation platforms. Following the same protocol as in our main experiments, we reformulated MedDG cases into simulated multi-agent medical consultation scenarios, in which the Patient Agent reveals symptoms only in response to targeted inquiries from the Doctor Agent.

Since MedDG predominantly covers common diseases, such as "cold", rather than specific diseases, such as "acute bronchitis", strict correctness judgments are less reliable than on MedQA and CMB. As a result, the accuracy differences between MedKGI and competing methods are relatively small. However, MedKGI still requires the fewest dialogue turns to a diagnosis, indicating that our information-gain-based symptom selection continues to deliver efficiency gains even in common-disease diagnosis.

Table 2: Supplementary results on MedDG.

Baselines	Rounds	Acc (%)
AgentClinic	14.42	82.17
Medical-CoT	16.64	81.19
CoD	12.81	**87.13**
STF-GT	11.88	82.17
Ours	**8.43**	85.15

B.4 Ablation Experiments

We tested three variants for the ablation study: (1) removing KG integration; (2) disabling the Clinical Record module; and (3) replacing the information gain-based symptom pruning strategy with random or frequency-based alternatives. As shown in Table 3, the full framework consistently achieves the highest diagnostic accuracy. Removing KG integration leads to a significant performance drop, underscoring the critical role of structured external knowledge. Meanwhile, omitting dialogue history summarization results in incomplete patient records, which impairs contextual coherence over multi-turn interactions. Finally, both random and frequency-based symptom pruning strategies result in lower accuracy than our information-gain approach, confirming that targeted, discriminative symptom selection is essential. Together, these findings validate the necessity of each key component in our design.

Table 3: Ablation experiments on AgentMedQA using Qwen3-8B, demonstrating the contribution of each component to diagnostic performance.

Method	Rounds	Acc (%)
w/o Knowledge Graph	13.44	44.34
w/o Clinical Record	12.09	57.55
Random node selection	19.25	31.13
Degree-based node selection	17.47	47.17
Ours	**9.11**	**69.81**

B.5 Entity Alignment Evaluation

Setup. To assess the quality of our entity alignment pipeline introduced in Section 4.2, we conducted a targeted evaluation on disease entity mapping. We sampled 100 disease names from the MedQA dataset and manually verified their alignment to corresponding PrimeKG nodes. Our results show that 76 out of 100 diseases were correctly mapped, indicating that the entity alignment pipeline successfully resolves the majority of disease entities, though some ambiguous or rare disease names remain challenging.

Results. Of the 100 sampled diseases, 76 were correctly mapped to their corresponding PrimeKG nodes, yielding an alignment accuracy of 76%. We further analyzed the 24 failure cases and identified the following error categories:

- Terminology divergence (19 / 24): Synonymous but lexically distinct terms that the embedding model failed to bridge (e.g., "Benign Paroxysmal Positional Vertigo" vs. "Benign Paroxysmal Positional Nystagmus").
- Missing entities (5 / 24): The disease is entirely absent from PrimeKG.

Impact on Downstream Performance. When an entity cannot be successfully mapped to the KG, the Bayesian posterior update is skipped for that entity. However, the remaining components of our framework continue to function normally, allowing MedKGI to fall back on the intrinsic parametric knowledge of the LLM for diagnostic reasoning.

B.6 Hyperparameter Selection

We performed controlled experiments to determine the optimal settings for two key hyperparameters. First, we examined how the number of candidate diseases affects accuracy. Figure 4(a) shows that accuracy peaks at 5 candidates and declines with more, as low-relevance candidates introduce noise. In practice, we recommend finding the optimal value by testing on a small sampled dataset. Second, we tested symptom sampling by defining k as the average symptoms per candidate disease. Figure 4(b) indicates optimal performance at $k = 1$, implying that focused symptom selection maximizes discrimination while avoiding redundancy.

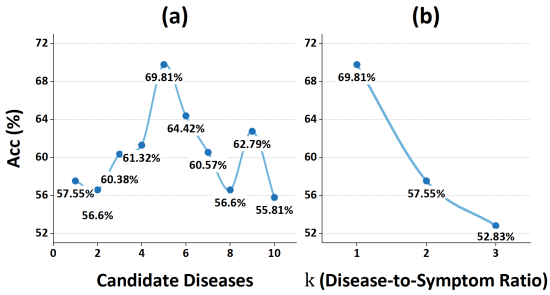

Figure 4: (a) Impact of candidate disease settings on accuracy. (b) Effect of k on Accuracy where k is defined as the ratio of the number of candidate diseases to the number of related symptoms from the KG.

B.7 Computational Efficiency and Latency Analysis

A potential concern for real-time diagnostic dialogues is the computational latency introduced by rebuilding the diagnostic subgraph and recalculating information gain at each turn. To assess this, we recorded the execution time of our framework during the multi-turn dialogue process. Since our framework retrieves symptoms directly associated with a set of candidate diseases, the size of the subgraph remains computationally manageable. Empirical measurements reveal that the dynamic subgraph construction takes an average of 0.57 ± 0.08 seconds per turn, while the information gain calculation requires approximately 6.13 ± 0.94 seconds. Furthermore, this latency is comparable to typical LLM API inference times (e.g., generating responses via GPT-4) (Benchmarks, 2026), ensuring that the overall user experience is not significantly degraded.

B.8 Qualitative Evaluation: Human and LLM-based Scoring

Beyond automated metrics, we conduct qualitative assessments from both LLM and human perspectives to evaluate dialogue quality, logical coherence, and clinical safety.

LLM-based Evaluation: We utilized GPT-4 as an impartial judge to automatically score the 100 generated dialogues across the three dimensions.

Expert Human Evaluation: To validate the LLM's judgments, we invited two board-certified medical professionals to conduct a blind manual evaluation on a random subset of 30 dialogues.

Table 4: Qualitative Evaluation Results (1-5 Scale).

Method	LLM Eval ($N = 100$)			Human Eval ($N = 30$)		
	Quality	Logic	Safety	Quality	Logic	Safety
AgentClinic	3.61	3.96	4.58	4.13	3.80	4.77
Ours	**4.62**	**4.01**	**4.64**	**4.38**	**3.95**	**4.80**

C Prompt Templates

C.1 Agent Initialization and Diagnosis Prompt

To simulate realistic clinical interactions, we implement a multi-agent diagnostic framework comprising three specialized agents: doctor agent, Patient Agent, and Measurement Agent with each guided by prompts to enforce specific behaviors and constraints.

The doctor agent adopts a constrained prompt specifying question limits T_{max} and tracks the count of asked questions $t_{current}$.

```
You are a doctor named Dr. Agent who only
responds in the form of dialogue. You are
inspecting a patient whom you will ask
questions in order to understand their
disease.
You are allowed to ask {T_max} questions
total before you must make a decision, and
have asked {t_current} questions so far.
```

Additionally, during the doctor agent 's differential diagnosis, we employ the following prompt template, which integrates patient demographics, recent dialogue history, structured clinical findings, and relevant medical knowledge extracted from the knowledge graph.

```
Age: {age}
Gender: {gender}
Chief Complaint: {chief complaint}

Recent dialogue history:
{3 recent dialogue turns}

Medical record
Chief Complaint: {chief complaint}
Symptoms: {symptoms}
Recent medical examinations:
{recent medical examinations}

Knowledge Graph Context
Relevant medical knowledge represented as
triples: {simplified subgraph triples}
```

The Patient Agent prevents patients from revealing diagnostic results directly, forcing the doctor agent to make diagnoses through symptom reasoning.

```
You're a patient in a clinic. The doctor
will ask questions or order exams to figure
out your illness.
Below is all of your information:
{patient profile}
Never name your disease and only describe
symptoms naturally: how they feel, when they
 flare, or how they affect you.
```

The Measurement Agent adopts a standardized result output format (`RESULTS: [results here]`), ensuring parseability of medical examination results.

```
You are a measurement reader who responds
with medical test results. Please respond in
 the format "RESULTS: [results here]".
Below is all of the information you have:
{medical examinations}. If the requested
results are not in your data, then you can
respond with NORMAL READINGS.
```

C.2 Diagnostic Record Initilization and Update Prompt

To ensure consistency of Diagnostic Record throughout the diagnosis process, we use a prompt that guides the LLM to perform evidence-based updates:

```
You are an experienced medical scribe.
Your task is to read the patient's latest
utterance and incrementally update the
structured summary below.
Rules:
    1. Add or revise only facts confirmed in
 the CURRENT utterance.
    2. Preserve all existing information
that is not contradicted.
    3. Use the exact JSON schema that was
provided (do not create new keys).
    4. Return only the updated JSON object,
with no extra commentary (without any
markdown formatting).

Schema: {schema}
Current structured summary:
{current diagnostic record}
Latest patient utterance: {schema}
```

C.3 Prompt for Entity Extraction and Evaluation

Symptom Entity Extraction. For symptom extraction from patient utterances, we employ the following prompt:

```
You are a helpful assistant with expertise
in medical symptom identification.
Please identify and extract all disease and
symptom entities from the following sentence.

Each entity must be no longer than 5
characters.
Rules:
    1. Include symptoms that are affirmed (
positive) in the "positive" list
    2. Include symptoms that are explicitly
denied (negative) in the "negative" list
    3. Pay special attention to negation
words like "no", "not", "don't", "haven't",
"can still", which typically indicate
negative symptoms
Return the result in valid JSON format as
shown below (without any markdown formatting
 and explanation):
    {
        "positive": ["Symptom 1", "Symptom
2", ...],
        "negative": ["Symptom 3", "Symptom
4", ...]
    }
Sentence: {sentence}.
```

Diagnostic Accuracy Judgment. To evaluate whether the Doctor Agent's final diagnosis matches the ground truth, we use the judgment prompt:

```
You are responsible for determining if the
correct diagnosis and the doctor's diagnosis
 are the same disease. Please respond only
with Yes or No. Nothing else.
Here is the correct diagnosis:
{ground truth diagnosis}
Here was the doctor diagnosis:
{doctor diagnosis}
Are these the same?
```

C.4 Prompt for OSCE Case Generation

We use the following prompt to generate standardized Objective Structured Clinical Examination (OSCE) scenarios for evaluation. The prompt instructs the LLM to produce a structured JSON containing patient demographics, symptom history, physical findings, test results, and the ground-truth diagnosis while providing only the clinical objective to the Doctor Agent.

```
Please generate a sample Objective
Structured Clinical Examination (OSCE) for
the patient actor and the doctor, including
what the correct diagnosis should be as a
structured JSON.

Only provide the doctor with the objective
and provide "test results" as a separate
category. Provide these for a primary care
doctor exam.

Generate an OSCE for the following case
study. Please read the answer category for
the correct diagnosis. Here is an example of
 correcting the OSCE format {example}.
Please create a new one here:
```

An OSCE-formatted Diagnostic Record is as follows:

```
{
  "OSCE Examination": {
    "Objective for Doctor": "Assess and diagnose the patient presenting with acute abdominal pain
.",
    "Patient Actor": {
        "Demographics": "30-year-old female",
        "History": "The patient complains of sudden onset of sharp, right lower quadrant
abdominal pain since last night. The pain has progressively worsened over the last 12 hours. She
mentions that she felt nauseous this morning but has not vomited. No recent changes in bowel
habits or urinary symptoms have been reported.",
        "Symptoms": {
            "Primary Symptom": "Sharp, right lower quadrant abdominal pain",
            "Secondary Symptoms": ["Nausea", "No vomiting", "No change in bowel habits", "No
urinary symptoms"]
        },
        "Past Medical History": "No significant past medical history. No previous surgeries.",
        "Social History": "Non-smoker, occasional alcohol use. Works as a software developer.",
        "Review of Systems": "Denies fever, vomiting, diarrhea, dysuria, or flank pain."
    },
    "Physical Examination Findings": {
        "Vital Signs": {
            "Temperature": "37.2 °C (99 °F)",
            "Blood Pressure": "120/75 mmHg",
            "Heart Rate": "78 bpm",
            "Respiratory Rate": "16 breaths/min"
        },
        "Abdominal Examination": {
            "Inspection": "No distension or visible masses.",
            "Auscultation": "Normal bowel sounds.",
            "Percussion": "Tympanic throughout, no shifting dullness.",
            "Palpation": "Tenderness in the right lower quadrant. No guarding or rebound
tenderness. Rovsing's sign positive, suggesting peritoneal irritation."
        }
    },
    "Test Results": {
        "Complete Blood Count": {
            "WBC": "12,000 /μL (elevated)",
            "Hemoglobin": "13.5 g/dL",
            "Platelets": "250,000 /μL"
        },
        "Urinalysis": {
            "Appearance": "Clear",
            "WBC": "2-5 /HPF",
            "RBC": "0-2 /HPF",
            "Nitrites": "Negative",
            "Leukocyte Esterase": "Negative"
        },
        "Imaging": {
            "Ultrasound Abdomen": {
            "Findings": "Enlarged appendix with wall thickening and fluid collection. No evidence
 of ovarian cyst or ectopic pregnancy."
            }
        }
    },
    "Correct Diagnosis": "Acute Appendicitis"
  }
}
```