ProbMedTOD: A Bayesian Network Guided Task-Oriented Dialogue System for Patient History Taking

Vishal Vivek Saley* IIT Delhi
Bhavesh Gurnani IIT Delhi
Dinesh Raghu IBM Research, India
Mausam IIT Delhi

Abstract

Patient history taking is a strategic, multi-turn process that refines diagnostic hypotheses through iterative questioning, mirroring clinical reasoning where each inquiry updates beliefs over candidate diagnoses. Deploying AI in this setting is constrained by privacy, cost, and latency, making Small Language Models (SLMs) a more desirable backbone compared to larger LLMs. However, SLMs capture only the fast, intuitive "System 1" side of this process and lack explicit mechanisms for the deliberative "System 2" reasoning needed to maintain and update diagnostic uncertainty over time. Bayesian Networks (BayesNets) offer a natural complement, providing an interpretable framework for auditable probabilistic belief tracking. Yet, clinical BayesNets are difficult to construct due to their reliance on expert-curated structure, dependencies, and parameters.

We introduce ProbMedTOD, which automatically synthesizes a clinical BayesNet from a small number of clinical notes in three stages: ontology building, structure prediction, and parameter estimation. ProbMedTOD then integrates the resulting BayesNet with SLMs, enabling probabilistic reasoning in multi-turn dialogues. A BayesNet Agent maintains real-time posterior distributions over diagnoses, while a Policy Agent picks diagnostically informative questions. Experiments on DDXPlus and MIMIC-IV Notes show that ProbMedTOD outperforms LLM-only and retrieval-based baselines, with especially strong gains for smaller models.

1 Introduction

Patient history taking is key to diagnosis [Engel and Morgan, 1973, Hampton et al., 1975, Boyd and Heritage, 2006]. It is a goal-driven, multi-turn dialogue to aimed identify the principal diagnosis, the condition mainly responsible for the visit [CMS, 2024]. Symptoms are often vague, partial, or overlapping, forcing clinicians to reason under uncertainty. Clinicians therefore use deliberative "System 2" reasoning, updating hypotheses and selecting queries by clinical value [Garibaldi and Olson, 2018]. Table 1 illustrates this: the doctor picks the query (Fatigue) to sharpen belief about TB.

Recently, large language models (LLMs) have shown strong results on medical QA and dialogue benchmarks [Li et al., 2023, Dou et al., 2023, Tu et al., 2024, Savage et al., 2024, Pal and Sankarasubbu, 2024, Singhal et al., 2025]. However, privacy and resource constraints in real clinical settings make SLMs an attractive practical choice. SLMs excel at the fast, fluent "System 1" side of dialogue, but lack an explicit mechanism to maintain and update a diagnostic belief or to select questions that maximize diagnostic value [Guan and Baral, 2021, Xu et al., 2024, He et al., 2024].

Bayesian Networks (BayesNets) have long been used for diagnosis under uncertainty [Shwe et al., 1991, Jaakkola and Jordan, 1999, Polotskaya et al., 2024], offering explicit belief updates and

*Corresponding author – vishal.vivek.saley@cse.iitd.ac.in

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: ProbMedTOD replicates deliberative diagnostic reasoning using a BayesNet.

Context: Patient reports a persistent **cough** for approximately three weeks. Initially, the cough presented with white to yellowish sputum (**colored/excess sputum**). Over the past week, the patient has developed **hemoptysis**, estimating the amount to be about half a teaspoon. The patient denies experiencing **dyspnea, chest pain, wheezing, rhinorrhea, or pharyngitis**. The patient does report **chest discomfort** associated with the cough.
Gold Response: How has your energy been? Any any fatigue?
Llama3 8B: Do you have a headache?
Llama3 8B + RAG: Have lost you sense of smell or taste? **Retrieved Documents:** 1. **Pertusis** causes coughing, loss of consciousness, whooping cough,.... 2. **Asthma** causes chest discomfort, colored/excess sputum, dyspnea, coughing,.... 3. **Bronchitis** causes chest discomfort, coughing, fatigue,....
ProbMedTOD: Do you have chills or fever? **Bayesian Inference:** 1. **Tuberculosis (89%)** causes coughing (82.7%), hemoptysis (36.3%), fever (76.5%),.... 2. **Bronchitis (5%)** causes chest discomfort (53.4%), coughing (86.9%), fatigue (49.2%),.... 3. **Pneumonia (5%)** causes coughing (90.4%), hemoptysis (90.4%), chills (67.9%)

transparent symptom-disease relationships. This raises a natural question: *can we combine the conversational strengths of SLMs with the reliability of Bayesian diagnosis?* To answer this question, we propose our framework ProbMedTOD, in which at each dialogue turn, a BayesNet Agent extracts symptom and risk-factor states and performs real-time Bayesian inference to update diagnostic posteriors. A Policy Agent uses these posteriors to select the most informative next question. Finally, a Diagnosis Agent predicts the top-5 principal diagnoses from accumulated evidence.

Unfortunately, clinical BayesNets are costly to build, requiring expert-curated ontologies, manually specified structures, and carefully estimated parameters. ProbMedTOD addresses this with an automated pipeline that constructs a BayesNet from a limited number of clinical notes in three stages: ontology construction identifies relevant diseases, symptoms, and risk factors; structure learning infers their dependencies; and parameter estimation assigns conditional probabilities directly predicted by an LLM—no expert annotation required.

In summary, our main contributions is ProbMedTOD, which to the best of our knowledge is the first system to combine BayesNets with SLMs for medical history taking, featuring an automated BayesNet construction pipeline and probabilistic multi-turn dialogue. On high-fidelity simulations on DDXPlus and MIMIC-IV Notes, ProbMedTOD significantly improves upon LLM-only and retrieval-based baselines, with especially strong gains for smaller models. Code and data are available at `https://github.com/dair-iitd/ProbMedTOD`.

2 Related Work

Our work broadly belongs to the line of research in which symbolic AI and/or task-specialized modules are combined with general-purpose language models for strong task performance in areas such as complex reasoning [Mittal et al., 2025], knowledge-base question answering [Patidar et al., 2024, Sawhney et al., 2025], relation extraction [Rathore et al., 2026], and dialog [Saley et al., 2024b]. Within the medical domain, neuro-symbolic approaches [Yang et al., 2025], such as those of Zhang et al. [2023] and Hong et al. [2024], extract logical rules from medical text to support symbolic reasoning. Related work by Zhang et al. [2025], Babakov et al. [2025] studies BayesNet structure discovery from ontology metadata. While similar in spirit, they typically do not perform probabilistic belief estimation or maintain a dialogue-conditioned belief distribution, capabilities central to ProbMedTOD.

With the rise of LLMs, their application in medical tasks has expanded rapidly, showing strong performance in medical QA and diagnosis benchmarks [Chen et al., 2023, Singhal et al., 2023, Savage et al., 2024, Pal and Sankarasubbu, 2024, Tu et al., 2024, Singhal et al., 2025, Sellergren et al., 2025]. Recent work has focused on adapting LLMs to medical dialogue settings. Li et al. [2023], Feng et al. [2026] fine-tuned LLaMA on doctor-patient conversations to build a QA system.

Other approaches, such as Xu et al. [2024] and He et al. [2024], use chain-of-thought prompting [Wei et al., 2022] to capture clinical reasoning, distilling this into smaller models. Several systems [Sree et al., 2024, Varshney et al., 2025, Rose et al., 2025] also integrate external medical knowledge through retrieval-augmented generation (RAG) [Lewis et al., 2020]; for example, Xu et al. [2024] uses RAG to infer likely diagnoses from dialogue context. However, none of these systems explicitly model diagnostic uncertainty. They rely on medical relationships learned during training, without probabilistic reasoning or belief updates. In contrast, ProbMedTOD maintains a diagnostic hypothesis using a BayesNet, enabling principled updates and structured decision-making under uncertainty.

BayesNets have a long history in medical diagnosis, offering a principled way to model uncertainty and reason under incomplete information. Seminal efforts include the probabilistic reformulation of the INTERNIST-1/QMR system [Shwe et al., 1991] and the use of variational inference for scalable reasoning in the QMR-DT network [Jaakkola and Jordan, 1999]. More recent work highlights the continued relevance of BayesNets in healthcare [Polotskaya et al., 2024].

Two closely related works are those by Guan and Baral [2021] and Pavez and Allende [2024]. Guan and Baral [2021] use a Bayesian framework to select the next symptom to inquire based on expected information gain, aiming to reduce diagnostic uncertainty. However, their system operates on structured inputs and lacks conversational capabilities. Pavez and Allende [2024] combine a BERT-based classifier with a Bayesian Network for mental health diagnosis, using BERT to generate a prior over diseases that is refined via probabilistic inference. While both approaches share our focus on uncertainty-aware reasoning, they do not leverage natural language dialogue. In contrast, ProbMedTOD integrates Bayesian inference with the fluent interaction abilities of LLMs, enabling data-driven, interpretable, and conversational medical history-taking.

3 Problem Definition

Let D be the set of m candidate diseases. A TOD system for medical history taking interacts with a patient over multiple turns. At each turn t, the system observes the dialogue history $H_t = \{(q_1, a_1), \ldots, (q_{t-1}, a_{t-1})\}$ of prior doctor and patient utterances denoted by q_i and a_i. It then selects the next doctor inquiry q_t. At the end of the dialogue, the system outputs the top-5 diagnoses.

A natural baseline is to prompt an LLM with H_t and D, asking it to decide the most informative q_t. This approach relies on implicit model reasoning and lacks an explicit mechanism to treat diagnostic uncertainty. In contrast, A BayesNet over risk factors R, symptoms S, and diseases D (Figure 3) provides a complementary, structured alternative. Given an evidence state consisting of confirmed and denied symptoms and risk factors $ES = (S^+, S^-, R^+, R^-)$, it computes a diagnostic believe as the posterior $DS = P(D \mid S^+, S^-, R^+, R^-)$, enabling explicit, interpretable belief updates.

In this work, we propose ProbMedTOD, the first TOD system that combines conversation capabilities of LLMs with probabilistic framework of BayesNets for medical history taking. A key challenge is enabling effective interaction between the LLM and the BayesNet during dialogue to extract evidence and update beliefs. Further, clinical BayesNets are difficult to construct, requiring expert-defined ontologies, graph structures, and parameters. We address this by automatically constructing a clinically valid BayesNet from a small corpus $\mathcal{P}$ of patient profiles.

4 ProbMedTOD System

We first describe the overall architecture of ProbMedTOD for medical history taking assuming that the BayesNet is given. Figure 1 provides a high-level overview of the system. ProbMedTOD consists of three main components: BayesNet Agent for evidence extraction and belief update, Policy Agent for dialogue policy, and Diagnosis Agent for final diagnosis. All three agents are prompt-based and use chain-of-thought (CoT) prompting.

BayesNet Agent: At turn t, BayesNet Agent processes the latest doctor-patient exchange (a_t, u_t) to update the evidence state ES_t and compute the diagnostic belief DS_t. A key challenge is that SLMs struggle to directly compare free-form patient utterances against large clinical ontologies. Ontologies often contain hundreds of expert terms, while patients describe symptoms in lay language (e.g., "can't catch my breath" $\rightarrow$ *dyspnea*). To bridge this gap, BayesNet Agent employs a retrieve–rerank pipeline that narrows the search space while leveraging the LLM's reasoning capabilities.

First, an LLM classifies whether u_t contains evidence about symptoms, risk factors, or neither. This lightweight routing step avoids unnecessary retrieval when the patient response is non-clinical (e.g., clarifications or pleasantries). Second, based on the classification, candidates are retrieved separately from the symptom set S and the risk-factor set R using the SapBERT [Liu et al., 2021] clinical retriever. The top-10 symptom nodes and top-10 risk-factor nodes are selected, skipping irrelevant retrievals as needed.

Finally, another LLM prompt resolves each candidate as confirmed, denied, or unmentioned, yielding per-turn evidences. These findings are aggregated into the cumulative evidence state ES_t. Finally, BayesNet Agent uses the BayesNet to compute the diagnostic belief state $DS_t = P(D \mid ES_t)$, a posterior distribution over diseases.

Policy Agent. The Policy Agent has two primary responsibilities: determining whether sufficient evidence has been gathered to make a diagnosis, and formulating the next inquiry to maximize diagnostic accuracy. To achieve this, the Policy Agent combines fast, heuristic-like System 1 reasoning with the principled, probabilistic beliefs of System 2 reasoning derived from the BayesNet.

Reasoning is crucial for the Policy Agent as it allows the agent to use information outside the BayesNet ontology and course-correct when the BayesNet belief DS_t is misaligned with the dialogue history H_t. The agent achieves this by prompting an LLM to reason through the evidence and dialogue history. If the Policy Agent concludes that enough evidence is available, it terminates the dialogue and invokes the Diagnosis Agent. Otherwise, it selects a follow-up question a_{t+1} targeting the leading source of diagnostic uncertainty in DS_t.

Diagnosis Agent When the policy concludes the interview at turn T, Diagnosis Agent receives the full history H_T and the final belief DS_T. Using the same reasoning pattern as Policy Agent, it weighs DS_T against the clinical evidence present in H_T, combining probabilistic and linguistic evidence to produce a ranked top-5 list of diagnoses from the candidate set $\mathcal{D}$.

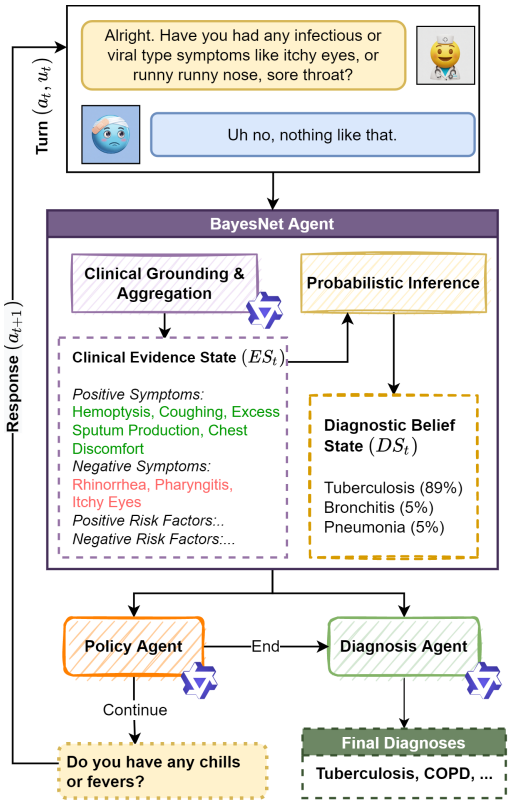


Figure 1: ProbMedTOD architecture. Extracted Evidence state is used to compute diagnostic belief via a BayesNet.

5 The BayesNet Construction

Clinical BayesNets are not always available, and constructing one from scratch is difficult. Defining the node set requires domain expertise because relevant symptoms and risk factors may be missing or inconsistently named across standard ontologies. Specifying the graph structure demands that clinicians reason explicitly about causal relationships among diseases, risk factors, and symptoms. Estimating the parameters is harder still: the strength of each relation must be quantified, but the patient-level data needed for statistical estimation are often costly to obtain. A BayesNet should also capture patterns in the target dataset, such as occupational or travel-related risk factors that may not appear in a standard ontology. To address these challenges, ProbMedTOD constructs a BayesNet from a small corpus $\mathcal{P}$ of patient notes for a given disease set D. Construction proceeds in three stages: *ontology building*, *structure prediction*, and *parameter estimation*. The notes identify the symptoms and risk factors that should be represented, medical references provide evidence for the graph structure, and an LLM maps the text in these sources to nodes, edges, and parameters.

Figure 3 shows the plate diagram for the BayesNet of our choice. The network contains three layers of binary nodes representing risk factors R, diseases D, and symptoms S. Risk factors include past

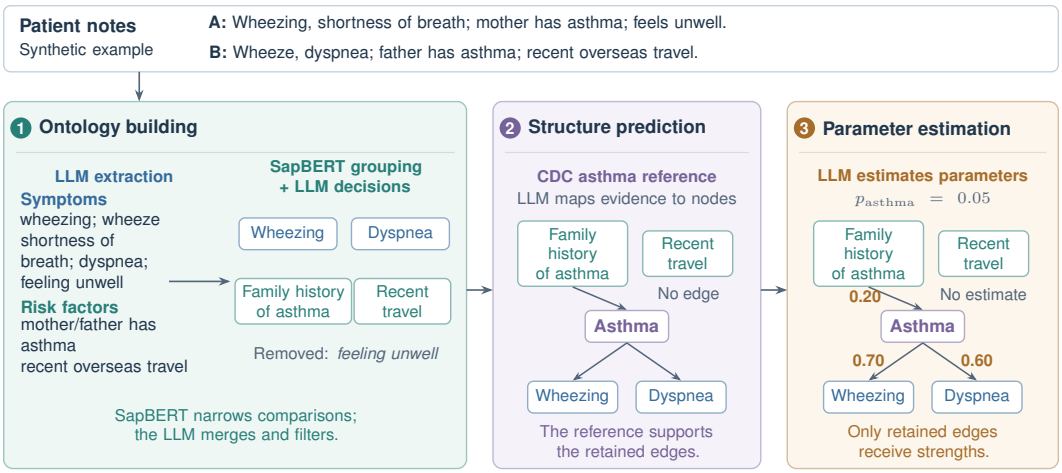


Figure 2: The three stages of BayesNet construction.

medical history and co-morbid conditions, family history, exposures, habits, and travel history. They modulate disease probabilities, while diseases generate observable symptoms. We parameterize both layers using Noisy-OR models [Jaakkola and Jordan, 1999]:

$$P(D_j{=}1 \mid R) = 1 - (1 - p_j) \prod_{k \in \mathrm{pa}(j)} (1 - r_{kj})^{R_k},$$

$$P(S_i{=}1 \mid D) = 1 - (1 - \lambda_i) \prod_{j \in \mathrm{pa}(i)} (1 - q_{ji})^{D_j}.$$

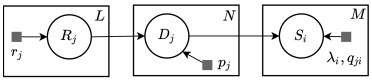


Figure 3: Plate diagram for the tripartite Bayesian network.

Here, $\mathrm{pa}(j)$ and $\mathrm{pa}(i)$ denote the parents of disease D_j and symptom S_i, respectively. The parameters r_{kj} and q_{ji} are the strengths of Risk→Disease and Disease→Symptom edges, p_j is the baseline probability of D_j, and λ_i is the symptom leak probability. We estimate p_j, r_{kj}, and q_{ji} with an LLM and fix every symptom leak to $\lambda_i = 0.001$.

At inference time, BayesNet Agent accumulates confirmed and denied clinical evidence $ES = (S^+, S^-, R^+, R^-)$ from the dialogue and computes disease posteriors $P(D_j{=}1 \mid ES)$. Following Guan and Baral [2021], we use a primary-diagnosis approximation in which one disease is active while the others are inactive:

$$P(D_j{=}1 \mid ES) \propto P(D_j{=}1 \mid R) \prod_{i \in S^+} P(S_i{=}1 \mid D_j) \prod_{i \in S^-} P(S_i{=}0 \mid D_j).$$

This formulation supports linear-time inference and real-time belief updates during dialogue. Appendix A provides the full derivation. We now describe the three construction stages.

Ontology Building. The first stage defines the node set from the clinical variables discussed in the patient notes $\mathcal{P}$. For each note, we prompt an LLM to extract symptoms and five types of risk factors: past medical history, family history, exposures, habits, and travel history. Pooling these mentions produces synonyms and vague terms. In Figure 2, for example, *wheezing* and *wheeze* refer to the same symptom, while *feeling unwell* is too vague to form a useful node.

Passing an entire category to the LLM would create long prompts and require it to compare many unrelated terms. We instead embed the terms using SapBERT [Liu et al., 2021] and use agglomerative clustering to form small groups of likely matches. We recursively divide clusters larger than 30 terms. The LLM then merges equivalent terms, assigns canonical names, and removes terms that are not useful for diagnosis. Thus, clustering narrows the candidate set, while the LLM makes the final decisions. We repeat this process for $k = 3$ rounds. The final symptom category defines S, and the other categories form the risk-factor set R.

Structure Prediction. Once the node set is fixed, we determine which nodes are connected. Co-occurrence in the small note corpus is not reliable evidence of a clinical relation, so we select a

medical reference for each disease. For every $D_j \in D$, the LLM generates search queries and fetches[2] candidate pages from authorative sources such as CDC, NHS and MedlinePlus. Once parsed, the LLM ranks these pages, and we select the highest-ranked page as the source document for D_j.

We predict edges separately for each disease-category pair, dividing each category into batches of at most 50 nodes. The LLM retains relations supported by the selected reference and orients them as $R_k \to D_j$ or $D_j \to S_i$. In Figure 2, recent travel remains in R because it occurs in the notes, but it has no edge to asthma because the reference does not support that relation. The union of the retained edges forms the final graph structure.

Parameter Estimation. With the network structure in place, we estimate the parameters of the Noisy-OR model. Estimating them from $\mathcal{P}$ would be unreliable because the corpus is small and is not intended to provide population-level frequencies. We therefore use the LLM to provide numerical estimates of the parameters. It estimates a baseline probability p_j for every disease, a strength r_{kj} for every retained Risk→Disease edge, and a strength q_{ji} for every retained Disease→Symptom edge. We do not estimate strengths for absent edges, and we fix every symptom leak to $\lambda_i = 0.001$. This produces a parameterized BayesNet that can be used by BayesNet Agent for the inference.

The pipeline can also use expert input when it is available. For example, an expert ontology, graph structure, or parameter set can replace the corresponding stage while the other stages remain unchanged. We study these replacements in our component ablation. We provide the prompts used for BayesNet construction in Appendix F.

6 Experimental Setup

We evaluate ProbMedTOD on principal diagnosis prediction using simulated multi-turn patient interviews. The system conducts a history-taking interview and returns a ranked list of candidate diagnoses. A system performs well when the true primary diagnosis appears near the top of this list.

Patient Simulator and Evaluation. Evaluating medical TOD systems is challenging because several valid questioning paths can lead to the same diagnosis. Existing turn-level protocols, such as action F1 and BLEU [Budzianowski et al., 2018, Hosseini-Asl et al., 2020, Saley et al., 2024a], compare each system turn with a fixed reference dialogue. This is too restrictive for clinical settings. For example, asking about fever before chills can be just as reasonable as asking in the opposite order. These metrics can therefore penalize clinically appropriate behavior and fail to capture the flexibility needed for diagnostic reasoning.

We instead use a dialogue-level evaluation framework. A patient simulator [Luo et al., 2024, Liu et al., 2025, Feng et al., 2026, Kyung et al., 2025] interacts with ProbMedTOD over multiple turns. The simulator is powered by Qwen3.8-27B [Qwen Team, 2026b] and generates patient responses conditioned on the underlying clinical note. This setup tests whether BayesNet-guided question selection helps ProbMedTOD elicit more diagnostically useful information than unstructured LLM querying. The full simulator prompt is provided in section F.8. Our standard simulator answers negatively when a queried fact is absent from the clinical note. Because real patients may instead guess, be unsure, or respond ambiguously, we also consider a noisy simulator that randomly gives a plausible diagnosis-consistent answer, reports the fact as unavailable, or expresses uncertainty.

At the end of each simulated interview, ProbMedTOD outputs a ranked list of candidate diagnoses. We compare this list with the true primary diagnosis, encoded using the first three characters of the ICD code. We report Mean Reciprocal Rank (MRR) that gives partial credit when the correct diagnosis is ranked highly and Hit@k ($k = 1, 3, 5$) that measures the proportion of cases where the correct diagnosis appears in the top k predictions. Together, these metrics evaluate diagnostic utility at the dialogue level, without penalizing valid alternative questioning strategies.

Datasets. We conduct experiments on DDXPlus and two specialities from MIMIC-IV Notes. DDXPlus [Fansi Tchango et al., 2022] is a dataset of synthetic patient profiles generated from a proprietary medical knowledge base. It covers 49 conditions across multiple specialties. Each profile contains a structured set of symptoms and antecedents, together with a differential diagnosis list.

[2]We use Tavily API for web searches.

We use 144 cases to construct our BayesNet and 882 cases for evaluation. In addition, we construct **BayesNet (Oracle)** from the one million patient cases available with DDXPlus. This data-driven network serves as an upper bound on the BayesNet constructed from the small note corpus.

MIMIC-IV Notes[Johnson et al., 2023] contains de-identified discharge notes from real hospital encounters. As a representative case, we focus on pulmonary diseases, a specialty with clear ICD-coded primary diagnoses, including pneumonia, tuberculosis, and heart failure. Following Kyung et al. [2025], we use the notes from patients' first visits and map ICD-9-CM codes to ICD-10-CM, retaining only unambiguous one-to-one mappings. We then remove diagnoses with nonspecific qualifiers (e.g., "unspecified," "unidentified," or "defined elsewhere"). Finally, we use Gemini-3.8-Flash to only retain cases where the diagnosis ICD-10-CM code strongly agrees with the discharge diagnosis from the notes. The resulting *pulmonary* subset contains 17 diseases at the three-character ICD-10-CM level, with 66 notes for BayesNet construction and 197 cases for evaluation.

Finally, we include a second MIMIC-IV Notes specialty, *abdominal*. We use the dataset released by Hager et al. [2024] consisting of four common diseases. We use 100 cases for BayesNet construction and 599 cases for evaluation. Dataset statistics are summarized in table 12.

Baselines. Our goal is to test whether ProbMedTOD improves the diagnostic performance of SLMs ($\leq$8B parameters). We evaluate three models: Nemotron-3-Nano-4B (Nano-4B) [NVIDIA Corporation, 2026], Qwen3-4B, and Qwen3-8B. We compare four configurations for each SLM model. **Base** performs history taking and diagnosis only using model's capabilities. **Base + RAG** augments the model with retrieved medical evidence. **Base + BayesNet** uses the automatically constructed network for belief updates and question selection. **Base + RAG + BayesNet** combines both forms of augmentation. On DDXPlus, we additionally evaluate **Base + BayesNet (Oracle)** to assess the scope of improvement when a large patient corpus available for BayesNet construction.

For RAG baselines, we use the set of document fetched during BayesNet construction. We break down the documents into chunks and retrieve the top-10 most similar chunks based on language model generated queries. These chunks are added to the model context for generating the next inquiry or the final diagnosis, without blowing up the context size. We provide further details in section C.

We evaluate large models, Qwen3-32B [Qwen Team, 2025], Qwen3.5-122B [Qwen Team, 2026a] and Gemini-3.8-Flash [Google DeepMind, 2026] under Base and Base + BayesNet configurations.

7 Results

We organize the evaluation around two research questions: a) how much does BayesNet guidance improve diagnostic performance for ProbMedTOD, and where do the gains come from? and b) how effective and reliable is the automatically constructed BayesNet itself?

7.1 ProbMedTOD Performance and Sources of Gains

ProbMedTOD improves performance for SLMs. Table 2 shows that ProbMedTOD variants improve performance across all metrics for SLMs on DDXPlus and MIMIC-IV (Pulmonary) datasets. MRR rises by 0.09–0.11 for Qwen3 4B and Qwen3 8B, and by 0.04–0.06 for Nano 4B. The consistence gains across two datasets showcase benefit of ProbMedTOD across various clinical settings. BayesNet guidance also complements RAG. Relative to the RAG baseline, ProbMedTOD improves MRR in seven of eight model–dataset comparisons and ties in the remaining one. Notably, when paired with larger models, Qwen3 32B, Qwen3.5 122B and Gemini-3.8-Flash, ProbMedTOD shows little improvement. We conduct a detailed analysis in section B.5 to understand why the larger models benefit less from BayesNet guidance. The results suggest that larger models already perform much of the required diagnostic reasoning implicitly, leaving limited margin for explicit BayesNet guidance. We also evaluate ProbMedTOD on the MIMIC-IV (Abdominal) dataset and observe that ProbMedTOD improves performance across SLMs. Results are reported in table 9.

Section E presents qualitative examples of ProbMedTOD in action. Given the patient's chief complaint, the base model overlooks the correct diagnosis and pursues less relevant lines of questioning. Consequently, the final diagnosis is incorrect. In contrast, the ProbMedTOD-augmented model incorporates the correct diagnosis into its diagnostic belief state early in the interview. This guides

Table 2: Results by models across datasets with standard patient simulator. **Bold** = the best scores withing group. *-results significantly better than paired baseline ($p < 0.05$, McnemarTest for Hit@k, WilcoxonTest for MRR).

Model	DDXPlus				MIMIC-IV (Pulmonary)			
	MRR	Hit@1	Hit@3	Hit@5	MRR	Hit@1	Hit@3	Hit@5
Nano 4B	0.27	0.21	0.32	0.36	0.31	0.17	0.44	0.56
+ BayesNet	**0.31***	**0.23**	**0.38***	**0.44***	**0.37**	**0.22**	**0.49**	**0.61**
Nano 4B + RAG	0.26	0.20	0.31	0.36	0.32	0.20	0.43	0.53
+ BayesNet	**0.36***	**0.28***	**0.43***	**0.49***	**0.41***	**0.27**	**0.50**	**0.64***
Qwen3 4B	0.45	0.37	0.53	0.58	0.46	0.36	0.55	0.61
+ BayesNet	**0.55***	**0.47***	**0.63***	**0.68***	**0.55***	**0.45***	**0.63***	**0.71***
Qwen3 4B + RAG	0.49	0.41	0.56	0.59	0.48	0.36	0.59	0.69
+ BayesNet	**0.56***	**0.47***	**0.64***	**0.69***	**0.56***	**0.45***	**0.65**	**0.71**
Qwen3 8B	0.47	0.37	0.57	0.63	0.49	0.38	0.59	0.68
+ BayesNet	**0.56***	**0.48***	**0.63***	**0.67***	**0.60***	**0.49***	**0.67***	**0.77***
Qwen3 8B + RAG	0.49	0.41	0.55	0.60	**0.57**	**0.47**	**0.68**	0.71
+ BayesNet	**0.57***	**0.49***	**0.65***	**0.71***	**0.57**	0.46	0.66	**0.74**
Qwen3 32B	**0.61**	0.52	**0.70**	**0.75**	**0.58**	0.46	**0.69**	**0.80**
+ BayesNet	**0.61**	**0.53**	0.69	0.74	**0.58**	**0.48**	0.67	0.76
Qwen3 32B + RAG	0.51	0.44	0.58	0.63	0.56	0.41	**0.71**	**0.78**
+ BayesNet	**0.60***	**0.51***	**0.69***	**0.73***	**0.62***	**0.53***	0.70	0.77
Qwen3.6 122B	**0.75**	**0.66**	**0.84**	**0.88**	0.69	0.58	0.80	**0.87**
+ BayesNet	**0.75**	**0.66**	**0.84**	0.87	**0.71**	**0.60**	**0.82**	**0.87**
Gemini 3.8 Flash	**0.78**	**0.70**	**0.85**	**0.89**	**0.74**	0.63	**0.84**	**0.92***
+ BayesNet	0.76	0.67	**0.85**	**0.89**	**0.74**	**0.64**	0.81	0.88

more systematic questioning for confirming/ruling out the diagnostic hypotheses, leading to a accurate final diagnosis.

ProbMedTOD remains robust under noisy evidence. While earlier simulations assume clean patient responses, real-world clinical encounters are often incomplete or ambiguous. We thus evaluate ProbMedTOD using a noisy patient simulator that introduces guessing, uncertainty, and uninformative responses. Although evidence noise lowers absolute performance overall, table 7 demonstrates consistent relative gains. ProbMedTOD improves MRR across all six SLM comparisons by 0.02–0.07, alongside gains or ties across Hit@k metrics. Results for larger models remain mixed. These findings confirm that the performance gains for SLMs are sustained under noisy conditions.

Both question steering and final diagnosis contribute to gains. To isolate whether improvements stem from better evidence gathering during the interview or from integrating the final posterior at diagnosis, we ablate ProbMedTOD incrementally. As shown in table 3, adding BayesNet guidance solely for question steering (Policy Agent) improves MRR across all four SLM–dataset settings by 0.01–0.06. Additionally providing the BayesNet posterior at diagnosis time (Diagnosis Agent) yields a further MRR increase of 0.04–0.08 across models. Across all configurations, the complete pipeline achieves the highest performance on every metric, raising overall MRR by 0.09–0.11 over the Base model and confirming that both question steering and posterior-guided diagnosis contribute to the final gain. We report futher studies regarding comorbidity status and alternative question-selection strategies in sections B.2 and B.4.

7.2 Evaluation of the Constructed BayesNet

The gains from BayesNet guidance raise a further question: how well does the automatically constructed network capture the diagnostic relationships that ProbMedTOD relies on? We first assess the network's diagnostic performance and posterior confidence independently of the language models, then use oracle comparisons to examine the quality of its parameters and locate the remaining limitations in construction.

Table 3: Effect of BayesNet-guided question steering and diagnosis. **Bold** marks the best score within each model and dataset.

Model	**DDXPlus** ($n = 882$)				**MIMIC-IV (Pulmonary)** ($n = 197$)			
	MRR	**Hit@1**	**Hit@3**	**Hit@5**	**MRR**	**Hit@1**	**Hit@3**	**Hit@5**
Qwen3 4B	0.45	0.37	0.53	0.58	0.46	0.36	0.55	0.61
+ BayesNet Steering	0.51	0.42	0.58	0.63	0.50	0.37	0.60	0.70
+ BayesNet Diagnosis	**0.55**	**0.47**	**0.63**	**0.68**	**0.55**	**0.45**	**0.63**	**0.71**
Qwen3 8B	0.47	0.37	0.57	0.63	0.49	0.38	0.59	0.68
+ BayesNet Steering	0.48	0.38	0.59	0.65	0.53	0.43	0.62	0.73
+ BayesNet Diagnosis	**0.56**	**0.48**	**0.63**	**0.67**	**0.60**	**0.49**	**0.67**	**0.77**

Table 4: Diagnostic performance from direct BayesNet inference on test-set.

Dataset	Method	N	MRR	Hit@1	Hit@3	Hit@5
DDXPlus	Uniform random	882	0.09	0.02	0.06	0.10
DDXPlus	BayesNet (Gemini-3.8-Flash)	882	**0.53**	**0.38**	**0.64**	**0.73**
MIMIC-Pulm	Uniform random	197	0.19	0.06	0.17	0.28
MIMIC-Pulm	BayesNet (Gemini-3.8-Flash)	197	**0.45**	**0.27**	**0.52**	**0.70**

The constructed network provides useful diagnostic signal. Given test-set evidence, we rank diseases by their BayesNet posterior probabilities and compare against uniform random ranking over the same disease set. Table 4 shows the results. On DDXPlus, the BayesNet achieves an MRR of 0.53 and Hit@5 of 0.73, compared with 0.09 and 0.10 for random ranking. On MIMIC-Pulm, the corresponding scores are 0.45 and 0.70, compared with 0.19 and 0.28. The network therefore captures diagnostic relationships that are informative even without the language-model agents.

Posterior Calibration. For these diagnostic rankings to support probabilistic decision-making, the confidence assigned to them also matters. To evaluate this confidence, we conduct calibration study using Brier@k, comparing the total probability assigned to the top-k diagnoses with whether the correct diagnosis appears in that set. The raw posteriors tend to be overconfident, with Brier@1 scores of 0.39 on DDXPlus and 0.44 on MIMIC-Pulm (table 5). Importantly, fitting a single temperature parameter on each dataset's training split reduces test-set Brier@1 to 0.22 and 0.18, respectively, with improvements also in Brier@3 and Brier@5. The network's diagnostic signal thus comes with confidence estimates that can be improved through a simple calibration step.

Oracle comparisons support the disease–symptom estimates. These predictive results establish the network's utility, but leave open how closely its underlying probabilities reflect empirical diagnostic relationships. We compare its parameters with those of the oracle BayesNet constructed from one million DDXPlus cases (section 6). For each parameter, we compute the KL divergence between the oracle and estimated probabilities, then average within each parameter category; lower values indicate closer agreement. Two sampling baselines draw probabilities uniformly at random or from a fitted Gaussian distribution, with results reported over 100 seeded samples. For disease–symptom probabilities, the constructed BayesNet achieves an average KL of 0.38, compared with 0.50 for uniform sampling and 0.41 for Gaussian sampling. This agreement supports the quality of the disease–symptom estimates. However, disease priors and risk factor–disease probabilities are not directly comparable across the two networks: the constructed network estimates them for the general population, including healthy individuals, whereas the oracle estimates them for a population consisting only of patients with a disease. Their KL values, and consequently the overall average, cannot be interpreted directly as estimation accuracy.

Structure refinement offers the largest gain in the oracle ablation. Parameter agreement alone cannot establish which construction errors most affect downstream performance. To assess their impact, we progressively replace the synthesized ontology, structure, and parameters with their oracle counterparts while holding the downstream model fixed (table 6). Introducing the oracle ontology improves every metric for both Qwen models. The largest subsequent MRR increase comes from

Table 5: Test-set BayesNet calibration before and after temperature scaling, fitted on training splits ($T = 4.99$ for DDXPlus; $T = 6.85$ for MIMIC-Pulm). Brier@k measures cumulative top-k posterior mass (lower is better). Scaling improves calibration on both datasets, especially at top-1.

Dataset	Calibration	Brier@1	Brier@3	Brier@5
DDXPlus	Raw BayesNet	0.39	0.32	0.26
DDXPlus	Temperature-scaled	**0.22**	**0.22**	**0.19**
MIMIC-Pulm	Raw BayesNet	0.44	0.44	0.29
MIMIC-Pulm	Temperature-scaled	**0.18**	**0.24**	**0.20**

Table 6: BayesNet component ablations on DDXPlus. **Bold** = the best score within each model. * = statistically significant difference from the base ($p < 0.05$).

Model	**Configuration**	**MRR**	**Hit@1**	**Hit@3**	**Hit@5**
Qwen3 4B	Base	0.45	0.37	0.53	0.58
	+ Expert Ontology	0.54*	0.47*	0.60*	0.66*
	+ Expert Structure	0.67*	0.59*	0.74*	0.79*
	+ Expert Parameters	**0.70***	**0.62***	**0.78***	**0.82***
Qwen3 8B	Base	0.47	0.37	0.57	0.63
	+ Expert Ontology	0.54*	0.47*	0.61*	0.66*
	+ Expert Structure	0.66*	0.57*	0.74*	0.78*
	+ Expert Parameters	**0.71***	**0.63***	**0.78***	**0.82***

replacing the automatically predicted structure: 0.54 to 0.67 for Qwen3 4B and 0.54 to 0.66 for Qwen3 8B. Oracle parameters provide smaller additional increases to 0.70 and 0.71, respectively. Within this cumulative replacement sequence, structure refinement therefore offers the largest gain, making it a promising target for improving automatic construction.

The constructed BayesNet thus supplies useful diagnostic guidance despite imperfections in its construction: its rankings are informative, temperature scaling improves its posterior confidence, and its disease–symptom estimates agree more closely with the oracle than the sampling baselines do. The oracle ablation points to graph structure as a promising route to further downstream gains.

8 Conclusion

We propose ProbMedTOD, the first TOD system for medical history taking that integrates SLM's conversational capabilities with interpretable probabilistic reasoning of BayesNets. ProbMedTOD leverages a dedicated BayesNet Agent that maintains running diagnostic beliefs over candidate diagnoses and a Policy Agent selects clinically informative questions, thus combining System 1 and System 2 reasoning in a unified framework. To avoid costly curation of BayesNets, ProbMedTOD proposes an automated pipeline that sythesizes BayesNet ontology, structure, and parameters based on given set of diseases and small corpus of patient profiles. The pipeline is flexible, allowing expert intervention at any stage, for further improvements.

We validate ProbMedTOD on two datasets: DDXPlus and MIMIC-IV covering different clinical specialties. Results show that ProbMedTOD consistently improves performance of the paired base models across the metrics, with larger gains for smaller models. Ablation studies confirm importance of each pipeline component, with graph structure learning being the most impactful. BayesNet guidance also leads to more clinically relevant questions and more accurate diagnoses in qualitative analysis. Together, these results highlight the benefits of combining structured Bayesian inference with conversational LLMs, and point toward a more transparent, reliable, and clinically grounded paradigm for medical dialogue systems.

Acknowledgments

Mausam is supported by research grants from Google, IBM AI Horizon Networks (AIHN), Wipro, and the Nick McKeown Chair Professorship. Vishal is supported from a Google Fellowship, NeuRIPS travel grant, and the IITD CSE department travel grant. We thank Microsoft AFMR program and Google Cloud Gemini Credits grant for access to LLM API calls. We thank the IIT Delhi HPC facility for its computational resources. This work was partially supported by the CSE Research Acceleration Fund of IIT Delhi

Limitations

While ProbMedTOD demonstrates strong performance across our evaluations, it has several limitations that warrant discussion. ProbMedTOD relies on a predefined set of diseases for constructing the Bayesian network. In practice, clinical settings serve highly diverse patient populations where the relevant disease space may shift over time or vary by region. Supporting such diversity would require automatic or on-the-fly extension of the Bayesian network. Although this direction is both interesting and technically feasible, we leave it to future work.

ProbMedTOD adopts a tripartite Bayesian network structure linking symptoms, diseases, and tests. While this structure is well-motivated for medical dialogue, it is only one of many possible graph topologies. Alternative structures, such as those incorporating risk factors, patient history, or treatment outcomes, may better capture the complexity of real clinical reasoning and are worth exploring in future work.

The Bayesian network in ProbMedTOD is constructed using Gemini 3.8 Flash, a relatively lightweight model chosen for efficiency. Using larger, more capable LLMs for this construction step may yield higher-quality conditional probability estimates and richer graph structures, potentially improving downstream dialogue performance.

Our experiments are conducted exclusively in English. Medical dialogue systems must ultimately serve multilingual populations. Wether the probabilistic reasoning encoded in the Bayesian network transfers effectively to other languages or clinical terminologies is an interesting question.

Impact Statement

ProbMedTOD advances development of AI systems for medical diagnosis, with potential to improve patient outcomes and reduce healthcare costs. By integrating probabilistic reasoning through Bayesian networks, our approach enhances the accuracy and reliability of diagnostic suggestions in simulated clinical dialogues. This could lead to more informed decision-making by clinicians and better patient care.

Further, our method allows clinical practitioners to specify their own diagnostic knowledge as BayesNet components in structured form, which can be used to guide the dialogue and diagnosis. This is an interesting human-AI collaboration setting in medical settings, where clinicians can inject their expertise into the system to improve performance and trustworthiness.

Finally, ProbMedTOD works well with small language models, which helps democratize access to advanced diagnostic tools in resource-constrained settings, where large models may be impractical. By enabling effective medical dialogue systems with smaller models, our work has the potential to broaden the reach of AI-driven healthcare solutions and reduce disparities in access to quality care.

References

N. Babakov, E. Reiter, and A. B. Diz. Scalability of bayesian network structure elicitation with large language models: a novel methodology and comparative analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10685–10711, 2025.

E. Boyd and J. Heritage. Taking the history: Questioning during comprehensive history-taking. 2006.

P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gasic. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In

Proceedings of the 2018 conference on empirical methods in natural language processing, pages 5016–5026, 2018.

Z. Chen, A. Hernández-Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

N. CMS. Icd-10-cm official guidelines for coding and reporting fy 2025 – updated october 1, 2024 (october 1, 2024 – september 30, 2025). Technical report, U.S. Department of Health and Human Services, July 2024. URL `https://stacks.cdc.gov/view/cdc/158747`. Accessed: 2025-09-24.

C. Dou, Z. Jin, W. Jiao, H. Zhao, Z. Tao, and Y. Zhao. Plugmed: Improving specificity in patient-centered medical dialogue generation using in-context learning. *arXiv preprint arXiv:2305.11508*, 2023.

G. L. Engel and W. L. Morgan. Interviewing the patient. *(No Title)*, 1973.

A. Fansi Tchango, R. Goel, Z. Wen, J. Martel, and J. Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318, 2022.

Y. Feng, J. Wang, L. Zhou, Z. Lei, and Y. Li. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16952–16956. IEEE, 2026.

B. T. Garibaldi and A. P. Olson. The hypothesis-driven physical examination. *Medical Clinics*, 102 (3):433–442, 2018.

Google DeepMind. Gemini 3.8 flash model card. `https://deepmind.google/models/model-cards/gemini-3-8-flash/`, 2026.

H. Guan and C. Baral. A bayesian approach for medical inquiry and disease inference in automated differential diagnosis. *arXiv preprint arXiv:2110.08393*, 2021.

P. Hager, F. Jungmann, and D. Rueckert. MIMIC-IV-Ext Clinical Decision Making: A MIMIC-IV Derived Dataset for Evaluation of Large Language Models on the Task of Clinical Decision Making for Abdominal Pathologies. *PhysioNet*, May 2024. doi: 10.13026/2pfq-5b68. URL `https://doi.org/10.13026/2pfq-5b68`. Version 1.0.

J. R. Hampton, M. Harrison, J. R. Mitchell, J. S. Prichard, and C. Seymour. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. *Br Med J*, 2(5969):486–489, 1975.

Y. He, Y. Zhang, S. He, and J. Wan. Bp4er: Bootstrap prompting for explicit reasoning in medical dialogue generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2480–2492, 2024.

S. Hong, L. Xiao, X. Zhang, and J. Chen. Argmed-agents: Explainable clinical decision reasoning with llm disscusion via argumentation schemes. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5486–5493. IEEE, 2024.

E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191, 2020.

T. S. Jaakkola and M. I. Jordan. Variational probabilistic inference and the qmr-dt network. *Journal of artificial intelligence research*, 10:291–322, 1999.

A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

D. Kyung, H. Chung, S. Bae, J. Kim, J. H. Sohn, T. Kim, S. Kim, and E. Choi. PatientSim: A Persona-Driven Simulator for Realistic Doctor-Patient Interactions. *PhysioNet*, Oct. 2025. doi: 10.13026/vq0d-v871. URL https://doi.org/10.13026/vq0d-v871. Version 1.0.0.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.

F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online, June 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.naacl-main.334.

R. Liu, K. Xue, X. Zhang, and S. Zhang. Interactive evaluation for medical LLMs via task-oriented dialogue system. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.325/.

X. Luo, Z. Tang, J. Wang, and X. Zhang. DuetSim: Building user simulator with dual large language models for task-oriented dialogues. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5414–5424, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.481/.

C. Mittal, K. Kartik, Mausam, and P. Singla. Fcorebench: Can large language models solve challenging first-order combinatorial reasoning problems?, 2025. URL https://arxiv.org/abs/2402.02611.

NVIDIA Corporation. Nvidia-nemotron-3-nano-4b-bf16. https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-4B-BF16, 2026.

A. Pal and M. Sankarasubbu. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 21–46, 2024.

M. Patidar, R. Sawhney, A. K. Singh, B. Chatterjee, Mausam, and I. Bhattacharya. Few-shot transfer learning for knowledge base question answering: Fusing supervised models with in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9147–9165. Association for Computational Linguistics, 2024.

J. Pavez and H. Allende. A hybrid system based on bayesian networks and deep learning for explainable mental health diagnosis. *Applied Sciences*, 14(18):8283, 2024.

K. Polotskaya, C. S. Muñoz-Valencia, A. Rabasa, J. A. Quesada-Rico, D. Orozco-Beltrán, and X. Barber. Bayesian networks for the diagnosis and prognosis of diseases: A scoping review. *Machine Learning and Knowledge Extraction*, 6(2):1243–1262, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

V. K. Rathore, M. H. Faisal, P. Singla, and Mausam. Combining distantly supervised models with in context learning for monolingual and cross-lingual relation extraction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 45483–45509. Association for Computational Linguistics, 2026.

D. P. Rose, C.-C. Hung, M. Lepri, I. Alqassem, K. Gashteovski, and C. Lawrence. MEDDxAgent: A unified modular agent framework for explainable automatic differential diagnosis. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13803–13826, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.677. URL `https://aclanthology.org/2025.acl-long.677/`.

V. Saley, G. Saha, R. Das, D. Raghu, et al. Meditod: An english dialogue dataset for medical history taking with comprehensive annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, 2024a.

V. V. Saley, R. J. Das, D. Raghu, and Mausam. Synergizing in-context learning with hints for end-to-end task-oriented dialog systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5596–5612. Association for Computational Linguistics, 2024b.

T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.

R. Sawhney, S. Yadav, I. Bhattacharya, and Mausam. Iterative repair with weak verifiers for few-shot transfer in KBQA with unanswerability. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 24578–24596. Association for Computational Linguistics, 2025.

A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

M. A. Shwe, B. Middleton, D. E. Heckerman, M. Henrion, E. J. Horvitz, H. P. Lehmann, and G. F. Cooper. Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. *Methods of information in Medicine*, 30(04):241–255, 1991.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180, 2023.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

Y. B. Sree, A. Sathvik, D. S. H. Akshit, O. Kumar, and B. S. P. Rao. Retrieval-augmented generation based large language model chatbot for improving diagnosis for physical and mental health. In *2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, pages 1–8. IEEE, 2024.

T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.

D. Varshney, N. Behera, P. Katari, and A. Ekbal. Medprom: Bridging dialogue gaps in healthcare with knowledge-enhanced generative models. *ACM Transactions on Computing for Healthcare*, 2025.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

K. Xu, Y. Cheng, W. Hou, Q. Tan, and W. Li. 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, 2024.

X.-W. Yang, J.-J. Shao, L.-Z. Guo, B.-W. Zhang, Z. Zhou, L.-H. Jia, W.-Z. Dai, and Y.-F. Li. Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models. *arXiv preprint arXiv:2508.13678*, 2025.

H. Zhang, J. Li, Y. Wang, and Y. Song. Integrating automated knowledge extraction with large language models for explainable medical decision-making. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1710–1717, 2023. doi: 10.1109/BIBM58861.2023.10385557.

Y. Zhang, Y. Zhang, P. Kordjamshidi, and Z. Cui. Bayesian network structure discovery using large language models. *arXiv preprint arXiv:2511.00574*, 2025.

A Exact Bayesian Inference using Noisy-OR

Following [Guan and Baral, 2021, CMS, 2024], we adopt a primary diagnosis model, i.e., chiefly one disease is responsible for the patient's condition. Given a clinical evidence state $ES = \{\mathcal{S}^+, \mathcal{S}^-, \mathcal{R}^+, \mathcal{R}^-\}$ our goal is to compute $P(D_j = 1|ES)$ for each disease D_j. We start with marginalizing over all D_{j^-} combinations.

$$\begin{aligned} P(D_j = 1 \mid \mathcal{S}^+, \mathcal{S}^-, \mathcal{R}^+, \mathcal{R}^-) &= \sum_{\forall D_{j^-}} P(D_j = 1, D_{j^-} | \mathcal{S}^+, \mathcal{S}^-, \mathcal{R}^+, \mathcal{R}^-) \\ &= P(D_j = 1, D_{j^-} = 0 | \mathcal{S}^+, \mathcal{S}^-, \mathcal{R}^+, \mathcal{R}^-) \end{aligned}$$

Now, posterior is computed by applying Bayes' Rule.

$$\begin{aligned} &P(D_j = 1 \mid \mathcal{S}^+, \mathcal{S}^-, \mathcal{R}^+, \mathcal{R}^-) \\ &= \frac{P(\mathcal{S}^+\mathcal{S}^- | D_j = 1, D_{j^-} = 0) P(D_j = 1, D_{j^-} = 0, \mathcal{R}^+, \mathcal{R}^-)}{P(\mathcal{S}^+, \mathcal{S}^-)} \end{aligned}$$

Note that the term $P(\mathcal{S}^+, \mathcal{S}^-)$ is common for all diseases D_j. Applying conditional independence of symptoms given disease, we compute the posterior using the following equation.

$$\begin{aligned} P(D_j = 1 \mid \mathcal{S}^+, \mathcal{S}^-, \mathcal{R}^+, \mathcal{R}^-) \propto\ & P(D_j = 1, D_{j^-} = 0 \mid \mathcal{R}^+, \mathcal{R}^-) \\ & \times \prod_{i \in \mathcal{S}^+} P(S_i = 1 \mid D_j = 1, D_{j^-} = 0) \\ & \times \prod_{i \in \mathcal{S}^-} P(S_i = 0 \mid D_j = 1, D_{j^-} = 0) \end{aligned}$$

where $P(D_j = 1, D_{j^-} = 0 \mid \mathcal{R}^+, \mathcal{R}^-) = P(D_j = 1 \mid \mathcal{R}^+, \mathcal{R}^-) \prod_{\forall j' \neq j} (1 - P(D_{j'} = 1 \mid \mathcal{R}^+, \mathcal{R}^-))$ are disease probabilities give risk factor combinations. We again apply Noisy-OR model and compute

$$P(D_j = 1 \mid \mathcal{R}^+, \mathcal{R}^-) = 1 - (1 - p_j) * \prod_{k \in \text{parent}(D_j))} (1 - r_{kj})^{R_k}$$

where p_j is prior probability of disease D_j and r_{kj} is conditional likelihood of D_j given risk factor R_k

B Additional Results

B.1 Noisy Simulator Results

B.2 Results by Comorbidity Status

ProbMedTOD benefits SLMs with and without comorbidities. Table 8 stratifies the DDXPlus results by whether a patient has a comorbid condition. BayesNet guidance improves MRR for all

Table 7: Results by models across datasets with noisy patient simulator. **Bold** = the best scores withing group. *-results significantly better than paired baseline ($p < 0.05$, McnemarTest for Hit@k, WilcoxonTest for MRR).

Model	DDXPlus				MIMIC-IV (Pulmonary)			
	MRR	**Hit@1**	**Hit@3**	**Hit@5**	**MRR**	**Hit@1**	**Hit@3**	**Hit@5**
Nano 4B	0.24	0.18	0.29	0.34	0.28	0.16	0.36	0.49
+ BayesNet	**0.29***	**0.21**	**0.36***	**0.43***	**0.35***	**0.22**	**0.42**	**0.59***
Qwen3 4B	0.41	0.34	0.47	0.53	0.46	0.36	0.53	0.62
+ BayesNet	**0.43**	**0.35**	**0.48**	**0.56**	**0.52**	**0.41**	**0.62***	**0.69**
Qwen3 8B	0.44	0.34	0.52	0.59	0.49	0.37	0.58	**0.69**
+ BayesNet	**0.48***	**0.39***	**0.56***	**0.63**	**0.52**	**0.42**	**0.63**	**0.69**
Qwen3 32B	0.53	0.44	0.61	0.67	**0.61**	**0.50**	**0.72**	**0.78**
+ BayesNet	**0.55**	**0.46**	**0.62**	**0.68**	0.59	0.48	0.68	**0.78**
Qwen3.6 122B	**0.66**	**0.57**	**0.76**	**0.82**	**0.72**	**0.60**	**0.83**	**0.89**
+ BayesNet	**0.66**	0.56	**0.76**	0.80	0.69	0.57	0.81	0.88
Gemini 3.8 Flash	**0.71**	**0.62**	**0.80**	**0.85**	**0.73**	**0.62**	**0.85**	**0.91**
+ BayesNet	0.69	0.59	0.78	0.84	0.69	0.57	0.81	0.88

Table 8: MRR on DDXPlus, stratified by comorbidity status. **Bold** marks the higher score within each paired baseline–BayesNet comparison.

Model	DDXPlus		
	All (N=882)	**Comorbid** (N=387)	**No-comorbid** (N=495)
Nano 4B	0.27	0.29	0.24
+ BayesNet	**0.31**	**0.34**	**0.28**
Qwen3 4B	0.45	0.48	0.43
+ BayesNet	**0.55**	**0.59**	**0.53**
Qwen3 8B	0.47	0.49	0.45
+ BayesNet	**0.56**	**0.59**	**0.53**
Qwen3 32B	**0.61**	0.59	**0.63**
+ BayesNet	**0.61**	**0.62**	0.60
Gemini 3.8 Flash	**0.78**	0.74	**0.81**
+ BayesNet	0.76	**0.75**	0.77

three SLMs in both groups: by 0.05–0.11 on comorbid cases and by 0.04–0.10 on cases without comorbidities. Thus, the overall improvements are driven by both the patients with and without comorbidities. The larger models again exhibit limited headroom: their small gains on comorbid cases do not transfer to the no-comorbidity group, consistent with the scale-dependent trend in the main results.

B.3 Results on the MIMIC-IV (Abdominal) Dataset

Table 9 showcases performance of ProbMedTOD on MIMIC-IV (Abdominal) dataset. Since abdominal dataset only has four target pathologies, we adopt top-1 diagnostic accuracy as the evaluation metric. The results reported in Table 9 show trend similar to the DDXPlus and MIMIC-IV (Pulmonary) datasets. Notably, ProbMedTOD improves performance of Base models by and RAG settings for SLMs.

Guidance improves smaller models in the clean setting. HandHeld BayesNet improves top-1 accuracy for Nano 4B, Qwen3 4B, and Qwen3 8B both with and without RAG, with gains of 0.02–0.08. The largest gain is for Qwen3 4B with RAG, from 0.55 to 0.63. Results for larger models are mixed: Qwen3 32B ties its baseline at 0.75 but drops from 0.79 to 0.75 with RAG, Qwen3.6 122B improves from 0.76 to 0.77, and Gemini 3.8 Flash drops from 0.79 to 0.75.

Table 9: Results MIMIC-IV (Abdominal) dataset. **Bold** = the best scores withing group. .*-results significantly better than base model ($p < 0.05$ Mcnemar Test).

Model	Hit1
Nano 4B	0.43
Nano 4B + HandHeld BayesNet	**0.48***
Nano 4B + RAG	0.44
Nano 4B + RAG + HandHeld BayesNet	**0.48**
Qwen3 4B	0.58
Qwen3 4B + HandHeld BayesNet	**0.60**
Qwen3 4B + RAG	0.55
Qwen3 4B + RAG + HandHeld BayesNet	**0.63***
Qwen3 8B	0.52
Qwen3 8B + HandHeld BayesNet	**0.57***
Qwen3 8B + RAG	0.57
Qwen3 8B + RAG + HandHeld BayesNet	**0.61**
Qwen3 32B	**0.75**
Qwen3 32B + HandHeld BayesNet	**0.75**
Qwen3 32B + RAG	**0.79**
Qwen3 32B + RAG + HandHeld BayesNet	0.75*
Qwen3.6 122B	0.76
Qwen3.6 122B + HandHeld BayesNet	**0.77**
Gemini 3.8 Flash	**0.79**
Gemini 3.8 Flash + HandHeld BayesNet	0.75*

Benefits are less consistent in the noisy setting. Guidance improves Nano 4B from 0.45 to 0.47 and Qwen3.6 122B from 0.80 to 0.81, but reduces accuracy for the other four models. The largest decrease is for Qwen3 8B, from 0.64 to 0.58. Thus, the gains observed for smaller models in the clean abdominal setting do not consistently carry over to the noisy patient simulator.

B.4 Alternative Question-Selection Strategies

We compare ProbMedTOD with two alternative strategies. The Naive Bayes variant reuses the disease–symptom edges and parameter estimates constructed by ProbMedTOD, but adopts a simpler probabilistic structure. Information Gain instead greedily selects the question with the largest expected one-step reduction in diagnostic uncertainty.

The learned disease–symptom model is useful even in a simpler structure. As shown in table 10, Naive Bayes improves MRR over the Base model in all six SLM–dataset comparisons, by 0.02–0.07. On DDXPlus, its gains for Qwen3 4B and Qwen3 8B are significant across every metric. Because this variant uses the disease–symptom relations and estimates produced by our construction pipeline, its strong performance provides additional evidence that these learned components capture useful clinical dependencies. The complete BayesNet nevertheless achieves higher MRR than Naive Bayes in all six SLM settings, indicating that the full structure provides further value.

Greedy information gain is not aligned with final diagnostic accuracy. Information Gain trails the complete BayesNet in six of eight model–dataset comparisons and ties it in one, with Nano 4B on DDXPlus as the sole exception. For Qwen3 4B and Qwen3 8B, it yields little or no improvement over Base and is consistently below the complete BayesNet. This result is expected from its myopic objective: a question that maximally reduces uncertainty at the current turn need not produce the most useful sequence of evidence for the final diagnosis. In contrast, ProbMedTOD allows the policy model to interpret the BayesNet belief jointly with the dialogue history when selecting the next question.

Table 10: Comparison of question-selection structures on DDXPlus and MIMIC-IV (Pulmonary). **Bold** marks the best score within each model and dataset.

Model	Method	DDXPlus ($n = 882$)				MIMIC-IV (Pulmonary) ($n = 197$)			
		MRR	Hit@1	Hit@3	Hit@5	MRR	Hit@1	Hit@3	Hit@5
Nano 4B	Base	0.27	0.21	0.32	0.36	0.31	0.17	0.44	0.56
	Naive Bayes	0.29	0.22	0.36	0.42	0.33	0.21	0.45	0.54
	Information Gain	**0.39**	**0.30**	**0.48**	**0.53**	**0.37**	**0.24**	0.47	0.60
	BayesNet	0.31	0.23	0.38	0.44	**0.37**	0.22	**0.49**	**0.61**
Qwen3 4B	Base	0.45	0.37	0.53	0.58	0.46	0.36	0.55	0.61
	Naive Bayes	0.51	0.43	0.59	0.63	0.50	0.39	0.58	0.68
	Information Gain	0.46	0.37	0.54	0.60	0.45	0.31	0.56	0.69
	BayesNet	**0.55**	**0.47**	**0.63**	**0.68**	**0.55**	**0.45**	**0.63**	**0.71**
Qwen3 8B	Base	0.47	0.37	0.57	0.63	0.49	0.38	0.59	0.68
	Naive Bayes	0.54	0.46	0.61	**0.67**	0.52	0.44	0.58	0.66
	Information Gain	0.48	0.38	0.57	0.63	0.45	0.31	0.57	0.69
	BayesNet	**0.56**	**0.48**	**0.63**	**0.67**	**0.60**	**0.49**	**0.67**	**0.77**
Qwen3 32B	Base	**0.61**	0.52	**0.70**	**0.75**	0.58	0.46	**0.69**	**0.80**
	Naive Bayes	0.60	0.50	0.68	0.74	**0.60**	**0.51**	0.66	0.76
	Information Gain	0.51	0.41	0.61	0.67	0.54	0.39	0.66	0.79
	BayesNet	**0.61**	**0.53**	0.69	0.74	0.58	0.48	0.67	0.76

Table 11: MRR with and without BayesNet. Gold coverage denotes the percentage of cases where the final BayesNet differential includes the gold diagnosis. Changes are computed within gold-present and gold-absent groups.

Dataset	Model	Baseline MRR	With BayesNet	Gold Coverage (%)	Δ Gold Present	Δ Gold Absent
DDXPlus	Qwen3 4B	0.45	0.55	70.98	+0.21	-0.17
DDXPlus	Gemini 3.8 Flash	0.78	0.76	77.89	+0.05	-0.23
MIMIC-IV (Pulmonary)	Qwen3 4B	0.46	0.55	81.73	+0.14	-0.13
MIMIC-IV (Pulmonary)	Gemini 3.8 Flash	0.74	0.74	90.86	+0.01	-0.23

B.5 Diminishing Gains for Larger Models

We examine why Bayesian Network (BN) guidance benefits smaller models such as Qwen3 4B more than the stronger Gemini 3.8 Flash. To understand this difference, we compare baselines and corresponding ProbMedTOD predictions on sets of cases stratified by whether the gold diagnosis appears in the final BayesNet differential. Table 11 summarizes the results across DDXPlus cases and MIMIC-IV (Pulmonary) cases.

The results suggest that stronger baseline performance leaves less room for useful BayesNet guidance. When the BayesNet includes the correct diagnosis, Qwen3 4B gains 0.21 MRR on DDXPlus and 0.14 on MIMIC-IV (Pulmonary), whereas Gemini gains only 0.05 and 0.01, respectively. This pattern is consistent with the BayesNet supplying diagnostic candidates that the smaller model would otherwise miss or rank poorly, while providing information that the stronger model can already support on their own.

The reduced benefit does not coincide with lower BayesNet coverage. Gemini has higher gold coverage on both datasets. Nevertheless, excluding the correct diagnosis remains associated with losses, reaching 0.23 MRR for Gemini on each dataset. Thus, as baseline capability increases, the additional value of helpful guidance shrinks while the cost associated with misleading guidance persists. This balance explains the substantial net gains for Qwen3 4B and the absence of comparable gains for Gemini.

C Implementation Details

We run all experiments on up to 8 NVIDIA A100 80GB GPUs. We use vLLM for inference over base language models. Open-weight models, especially SLMs, can produce repetitive outputs or

Table 12: Dataset statistics for evaluation.

Dataset	Train # Cases	Test # Cases	# Diagnoses
DDXPlus	144	882	49
MIMIC-IV Notes (Pulmonary)	66	197	17
MIMIC-IV Notes (Abdominal)	100	599	4

Table 13: BayesNet Statistics

Dataset	DDXPlus	MIMIC-IV (Pulmonary)	MIMIC-IV (Abdominal)
#Disease	49	17	4
#Symptom	104	47	13
#Risk Factor	70	47	15
#Disease to Symptom	621	178	28
#Risk Factor to Disease	218	149	19
Construction Cost	$16.28	$7.94	$5.69

violate the required response format. To reduce these failures, we use vLLM structured generation with temperature set to 0.0, and our prompts explicitly forbid verbatim repetition. If a model violates the format or repeats itself, we resample up to five times with increasing temperature values: 0.1, 0.2, 0.3, 0.5, and 0.7. If no valid response is produced after five attempts, we halt the simulation. All our models run in bfloat16 precision without quantization. We run multiple simulations in parallel to reduce evaluation time.

For RAG baselines, we use the same set of document fetched during BayesNet construction. We divide these documents into overlapping chunks of 512 tokens. At each dialogue turn, the language model generates retrieval queries from the conversation so far. We embed the queries and chunks with the SapBERT (`cambridgeltl/SapBERT-from-PubMedBERT-fulltext`) [Liu et al., 2021] encoder and retrieve the 10 most similar chunks based on cosine similarity. These chunks are added to the model context for generating the next inquiry or the final diagnosis.

All our simulations take about 3 hours to run per model per dataset. BayesNet synthesis is a one time operation that takes about 4 hour per dataset. The majority of this time is spent on parameter estimation, which requires multiple calls to the LLM for each disease-symptom and risk-factor disease pair.

Note that DDXPlus is released under CC BY 4.0 license while MIMIC-IV is released under a custom license that requires credentialing and data use agreement. We have obtained necessary permissions to use these datasets for research purposes.

D Dataset and Bayesnet Statistics

Table 12 shows the statistics of the datasets used in our experiments. Table 13 shows the statistics of the Bayesnets used in our experiments, including the number of nodes, edges, and the average Markov blanket size.

E Qualitative Examples

E.1 PULM: Qwen3-8B, case 16

Case Summary. Patient has worsening shortness of breath for one day, preceded by three days of cough, congestion, low-grade fever, and night sweats. Increasing cough with mild hemoptysis, exertional dyspnea worse than baseline, and chest tightness temporarily relieved by inhalers. Similar episodes occur each winter.

Patient has a medical history of Bronchiectasis, prior right lower lobectomy, and reactive airway disease. Baseline exertional dyspnea limits walking around the house.

Gold Diagnosis: Bronchiectasis

The interviews below compare Qwen3-4B and Qwen3-4B+ProbMedTOD interacting with the simulated patient described above. Both models receive the same chief complaint: shortness of breath. Qwen3-4B begins by asking about the nature and duration of the breathing difficulty and associated symptoms. It subsequently focuses on travel history, immobility, and a family history of blood clots, incorrectly favoring pulmonary embolism. The model neither asks about prior lung disease or surgery nor identifies the gold diagnosis of bronchiectasis.

In contrast, Qwen3-4B+ProbMedTOD begins by asking about chronic cough and smoking history. Although bronchiectasis is absent from the initial Bayesian network belief state, the model quickly considers it based on chronic cough with blood-tinged sputum and the absence of a smoking history. It then systematically gathers evidence to assess competing diagnostic hypotheses. For example, it asks about fever and chills in turn 2 to assess tuberculosis and elicits a history of recurrent infections and lung disease to investigate bronchiectasis. This methodical history-taking leads to the gold diagnosis.

Chief Complaint: Doctor, I'm really struggling to catch my breath.

Base (13 doctor turns)

Dr (1): Can you describe the nature of your breathing difficulty?

Pt: It's really tight in my chest, and I feel like I can't get enough air. Even just getting out of bed or walking to the door makes me winded. I used to be able to walk around my house without stopping, but now I have to stop and catch my breath almost immediately.

Dr (2): How long have you been experiencing this difficulty breathing, and has it been getting worse over time?

Pt: It started about three days ago with a cold—cough, congestion, and a little fever. But the breathing really got bad just yesterday. It's been getting worse since then, not better.

Dr (3): Have you had any recent travel, or do you have a history of cardiovascular disease or smoking?

Pt: No, I haven't been traveling anywhere. I don't have any heart problems, and I don't smoke.

Dr (4): Have you been producing any sputum, and if so, what is its color, consistency, and volume?

ProbMedTOD (14 doctor turns)

Diagnostic Belief State: Asthma (37.71%); Heart failure (26.40%); Pneumonitis due to solids and liquids (13.23%); Bronchiectasis (8.35%); Emphysema (6.55%).

Dr (1): Have you had a chronic cough or a history of smoking?

Pt: I do have a cough, yes. It's been bothering me for a few days now, and I've been coughing up a little bit of blood. But no, I don't smoke. I haven't smoked in a long time.

Diagnostic Belief State: Bronchiectasis (65.56%); Pneumonitis due to solids and liquids (21.31%); Pulmonary edema (6.82%); Abscess of lung and mediastinum (3.42%); Respiratory tuberculosis (2.10%).

Dr (2): Have you had any fever or chills along with your cough?

Pt: Yes, I've had a low-grade fever. I've also been breaking out in night sweats, which is pretty unusual for me.

[Doctor turns 3–11 omitted]

Diagnostic Belief State: Bronchiectasis (57.42%); Pneumonitis due to solids and liquids (32.78%);

Pt: Yeah, I've been coughing up some stuff. It's not a lot, just a little bit, but it's got some dark blood in it. It's not a huge amount, just a few streaks or drops when I cough.

[Doctor turn 5 omitted]

Dr (6): Have you been immobile or bedridden for any period recently, or do you have a family history of blood clots or deep vein thrombosis?

Pt: No, I haven't been stuck in bed. I've been moving around, just slower than usual because of the breathing. As for family history, I don't know of anyone having blood clots or DVT. My family has had some cancers and heart issues, but not clots.

[Doctor turns 7–11 omitted]

Dr (12): Have you been on any long trips or had prolonged immobility recently that could increase your risk for blood clots?

Pt: No, I haven't been on any long trips. And I haven't been stuck in one place for a long time. I've been getting up and moving around, just not as much as I usually do because of the breathing.

Dr (13): Thank you for sharing all these details. I believe we have covered all the necessary information to proceed with the next steps. Let me review the findings and prepare for the next phase.

Final diagnosis turn

Reasoning: The patient presents with acute dyspnea that has progressively worsened over the past three days, accompanied by cough, congestion, low-grade fever, and night sweats. The presence of blood-tinged sputum, along with a history of no recent travel, immobility, or risk factors for pulmonary embolism, suggests that a pulmonary embolism may not be the primary concern. However, the progressive nature of dyspnea, nocturnal sweats, and fever raise a suspicion of an infectious process. The absence of a history of smoking or heart disease makes heart failure less likely, although it cannot be entirely ruled out without further evaluation. The presence of blood-tinged sputum and fever could point towards an infectious etiology, such as acute pharyngitis or tonsillitis, but the severity of respiratory symptoms suggests a more significant condition. The possibility of tuberculosis is noted given the presence of fever, night sweats, and cough, although the patient denies exposure. The presence of chest tightness and respiratory discomfort also raises the concern of acute bronchitis or even pneumonitis. Finally, the acute nature of symptoms, along with the absence of

Pulmonary edema (5.47%); Asthma (3.94%); Respiratory tuberculosis (0.23%).

Dr (12): Have you had a history of recurrent respiratory infections or a previous diagnosis of lung disease?

Pt: Yes, I have a lung condition called bronchiectasis. It's been with me for a long time. I actually had part of my right lung removed when I was a child because of it. I usually get these breathing problems once a year, mostly in the winter, but this time it feels a bit different.

Diagnostic Belief State: Bronchiectasis (57.42%); Pneumonitis due to solids and liquids (32.78%); Pulmonary edema (5.47%); Asthma (3.94%); Respiratory tuberculosis (0.23%).

Dr (13): Have you experienced any recent changes in the frequency or severity of your breathing difficulties, or have you had any episodes of chest pain or dizziness?

Pt: My breathing has definitely gotten worse. I used to be able to walk around my house before I had to stop, but now I get short of breath just getting out of bed or walking to the door. I don't have any chest pain, just that tightness, and I haven't felt dizzy at all.

Diagnostic Belief State: Bronchiectasis (63.77%); Pneumonitis due to solids and liquids (25.48%); Pulmonary edema (6.07%); Asthma (4.38%); Respiratory tuberculosis (0.15%).

Dr (14): Thank you for your time. I have gathered enough information to proceed with the next steps in your care.

Final diagnosis turn

Diagnostic Belief State: Bronchiectasis (63.77%); Pneumonitis due to solids and liquids (25.48%); Pulmonary edema (6.07%); Asthma (4.38%); Respiratory tuberculosis (0.15%).

Reasoning: The patient presents with dyspnea, cough with hemoptysis, low-grade fever, night sweats, chest tightness, and a history of bronchiectasis. These symptoms suggest chronic lung disease complications. The Bayesian Network Hints include Bronchiectasis, Asthma, and Pulmonary edema, which align with the patient's history. Given the patient's known bronchiectasis, this condition is a top priority. The intermittent fever, night sweats, and hemoptysis raise concern for infection, which could include tuberculosis. The wheezing and response to inhalers suggest asthma, though this may be secondary to the bronchiectasis. The patient's denial of leg swelling and weight gain makes pulmonary edema less likely. The history of

other risk factors, makes pulmonary embolism a possible differential, though the lack of immobility or other risk factors decreases its likelihood.

Differential:

1. Pulmonary embolism
2. Acute bronchitis
3. Acute pharyngitis
4. Respiratory tuberculosis
5. Pneumonitis due to solids and liquids

bronchiectasis and acute exacerbation supports its inclusion in the differential. Asthma is considered given the wheezing and inhaler use. Pulmonary edema is included as a possibility due to the respiratory symptoms, though the absence of peripheral edema makes it less likely. The presence of hemoptysis, fever, and night sweats supports respiratory tuberculosis as a possibility, though the patient denies exposure. Pneumonitis due to solids and liquids is included given the history of recent respiratory infection, though the absence of aspiration risk factors makes it less likely.

Differential:

1. Bronchiectasis
2. Asthma
3. Respiratory tuberculosis
4. Pulmonary edema
5. Pneumonitis due to solids and liquids

E.2 DDXPLUS: Qwen3-8B, case 1760

Case Summary. Patient is a 49-year-old woman with vomiting after coughing. Patient reports exposure to someone with pertussis (whooping cough). She reports being significantly overweight for her height. Vaccinations are up to date.

Gold Diagnosis: Whooping cough

The interviews below compare Qwen3-4B and Qwen3-4B+ProbMedTOD interacting with the simulated patient described above. Both models receive the same chief complaint: vomiting after coughing. Qwen3-4B begins by asking about the duration of the cough and accompanying symptoms. It then explores several diagnostic possibilities, asking about heartburn (GERD), recent chest trauma (spontaneous rib fracture), and stridor (croup). However, it neither considers whooping cough nor asks about exposure to sick individuals.

In contrast, Qwen3-4B+ProbMedTOD identifies whooping cough as a likely diagnosis in its Bayesian network–driven diagnostic belief state. Its subsequent questions focus on gathering evidence to distinguish among competing hypotheses. For example, it immediately asks about exposure to sick individuals, explores dietary changes to assess possible anaphylaxis, and asks about night sweats, fatigue, and weight loss to evaluate other systemic causes. This targeted questioning leads to the correct diagnosis.

Chief Complaint: I vomit after coughing.

Base (11 doctor turns)

Dr (1): How often have you been vomiting after coughing, and are there any other symptoms accompanying the vomiting?

Pt: It only happened once, actually. I don't have any other symptoms like a fever or headache, just the coughing and then the vomiting.

Dr (2): Have you ever had episodes of coughing that led to vomiting before, and do you have any known medical conditions such as asthma or chronic lung disease?

ProbMedTOD (11 doctor turns)

Diagnostic Belief State: whooping cough (55.77%); anaphylaxis (25.05%); pneumonia (6.78%); influenza (3.87%); chronic rhinosinusitis (2.68%).

Dr (1): Have you experienced any fever, sore throat, or difficulty breathing along with the vomiting after coughing?

Pt: No, I don't have a fever, my throat doesn't hurt, and I'm not having trouble breathing. It's just the coughing that makes me throw up.

Pt: No, I don't think I've had this happen before. And no, I don't have asthma or any lung problems that I know of.

Dr (3): Have you experienced any recent trauma or injury to your chest that might have caused a rib fracture?

Pt: No, I haven't had any injuries or trauma to my chest recently.

[Doctor turns 4–6 omitted]

Dr (7): Have you experienced any episodes of severe coughing or coughing fits recently that might have led to vomiting?

Pt: Yes, I've had some really bad coughing fits. They're pretty intense, and that's when I ended up throwing up.

Dr (8): Can you describe the cough—how long has it been going on, is it productive, and have you experienced any chest pain or discomfort along with it?

Pt: I don't have any chest pain or discomfort. I'm not sure exactly how long it's been going on, maybe a few days? And I don't think I'm coughing up any phlegm, it just feels like a dry, harsh cough.

[Doctor turn 9 omitted]

Dr (10): Have you experienced any heartburn, acid reflux, or nausea in addition to the coughing?

Pt: No, I don't have any heartburn or acid reflux. I didn't feel nauseous until right after the coughing fit, when I threw up.

Dr (11): Thank you for sharing your symptoms. Based on your history and the questions we've discussed, I believe we have gathered sufficient information to proceed with the next steps. Let's review what we've discussed and determine the best course of action.

Final diagnosis turn

Reasoning: The patient presents with a single episode of vomiting following intense coughing fits, with no prior history of similar events or underlying conditions like asthma or lung disease. The absence of chest trauma rules out rib fracture or pneumothorax. The lack of fever, headache, or other systemic symptoms suggests non-infectious causes. The cough is described as dry, harsh, and without productive sputum, pointing away from respiratory infections like bronchitis or pneumonia. The absence of stridor or respiratory distress rules out conditions like epiglottitis or croup. The lack

Diagnostic Belief State: whooping cough (48.36%); anaphylaxis (32.10%); chronic rhinosinusitis (6.87%); gerd (3.52%); pneumonia (2.03%).

Dr (2): Have you had any recent changes in your diet or exposure to sick individuals?

Pt: No, I haven't changed my diet. But I have been around someone who has whooping cough, so I'm worried I might have caught it.

Diagnostic Belief State: whooping cough (90.93%); anaphylaxis (5.63%); chronic rhinosinusitis (1.21%); gerd (0.62%); pneumonia (0.36%).

Dr (3): How long has the cough lasting, and have you noticed any episodes of a distinctive 'whoop' sound when breathing in?

Pt: I'm not sure exactly how long it's been going on, maybe a few days? And no, I haven't really noticed a specific 'whoop' sound, just the coughing fits that make me throw up.

[Doctor turn 4 omitted]

Diagnostic Belief State: whooping cough (94.94%); anaphylaxis (1.45%); chronic rhinosinusitis (1.27%); gerd (0.65%); pneumonia (0.38%).

Dr (5): Have you experienced post-tussive vomiting, where you vomit after coughing fits?

Pt: Yes, that's exactly what's happening. The coughing fits are so bad that I end up throwing up right after.

[Doctor turns 6–9 omitted]

Diagnostic Belief State: whooping cough (99.80%); pneumonia (0.15%); bronchitis (0.03%); inguinal hernia (0.01%); chronic rhinosinusitis (0.00%).

Dr (10): Have you experienced any recent weight loss, night sweats, or fatigue that could indicate a chronic or systemic infection?

Pt: No, I don't have any of those. I'm not losing weight, I don't sweat at night, and I don't feel tired or fatigued.

Diagnostic Belief State: whooping cough (99.93%); bronchitis (0.03%); pneumonia (0.03%); inguinal hernia (0.01%); anaphylaxis (0.00%).

Dr (11): Thank you for sharing all these details. Based on your symptoms and history, I have gathered sufficient information to proceed with the next steps in your care. Have a good day!

Final diagnosis turn

Diagnostic Belief State: whooping cough (99.93%); bronchitis (0.03%); pneumonia (0.03%); inguinal hernia (0.01%); anaphylaxis (0.00%).

of heartburn or acid reflux rules out GERD. Given the acute nature and the mechanism of injury (coughing leading to vomiting without respiratory compromise), the most likely diagnoses are spontaneous rib fracture (due to the sudden, intense coughing), bronchospasm/acute asthma exacerbation (as the cough could trigger bronchospasm despite no prior asthma history), bronchitis (as a dry cough may be a feature of viral bronchitis), pulmonary embolism (as sudden coughing fits could be associated with embolism), and acute laryngitis (due to the dry, harsh cough resembling acute laryngitis symptoms).

Differential:

1. spontaneous rib fracture
2. bronchospasm / acute asthma exacerbation
3. bronchitis
4. pulmonary embolism
5. acute laryngitis

Reasoning: The patient presents with severe coughing fits leading to post-tussive vomiting. The history includes exposure to whooping cough, but no 'whoop' sound, fever, fatigue, weight loss, or night sweats. The key features are coughing and vomiting without systemic symptoms. Based on the Bayesian hints and the history, the top five diagnoses are: whooping cough (exposure and symptoms consistent), bronchitis (coughing fits), acute laryngitis (coughing and vomiting), acute rhinosinusitis (coughing may be associated), and urti (common cold with coughing fits). The patient denies fever or systemic symptoms, which may lower the likelihood of other infections like pneumonia.

Differential:

1. whooping cough
2. bronchitis
3. acute laryngitis
4. acute rhinosinusitis
5. urti

F Prompts

F.1 Disease prior

```
<|system|> You are a medical epidemiology expert estimating population-level
disease prevalence.

Provide your best-calibrated estimate using general medical and epidemiological
knowledge.

Return the estimated number of current cases per 1,000,000 adults as an integer
from 0 to 1,000,000. This representation is used to preserve precision for rare
diseases. Do not round the estimate to a percentage or automatically return zero
merely because the disease is rare.

<|user|>
Estimate the point prevalence of {disease_name} in the general adult population.

Interpret point prevalence as the number of randomly selected adults per
1,000,000 who currently have the disease.
```

F.2 Symptom given disease

```
<|system|> You are a medical expert estimating conditional symptom frequencies
for a coarse-grained clinical ontology.

Estimate P(symptom | disease): the probability that a typical adult with the
disease has the symptom during the active disease episode and would acknowledge
it if specifically asked during an initial clinical evaluation.

Interpretation rules:
- Estimate symptom presence, not the probability that it is mentioned
spontaneously.
- Consider the general adult patient population across typical clinical
settings.
```

```
- If the disease label is broad, consider its standard clinical scope and
average across its common subtypes and presentations according to their
approximate frequency. Do not restrict it to a single subtype.
- If the symptom label is broad, include clinically equivalent ways in which
that symptom may be experienced or described, but do not include unrelated
manifestations.
- Estimate the probability attributable to the current disease episode, not the
patient's lifetime probability of experiencing the symptom.
- Return a best-calibrated central estimate as an integer from 0 to 100.
- Use 0 or 100 only when the symptom is essentially never or almost always
present, respectively.
- Perform the clinical reasoning internally and do not explain it.

<|user|>
Estimate P(symptom | disease) for the following pair.

Disease: {disease_name}
Symptom: {symptom_name}

What percentage of typical adult patients with this disease have this symptom
during the active disease episode and would acknowledge it if specifically
asked?
```

F.3 Disease given risk factor

```
<|system|> You are a medical epidemiology expert estimating conditional disease
frequencies for a coarse-grained clinical ontology.

Estimate P(disease | risk factor): the probability that a randomly selected
adult with the specified risk factor currently has the disease.

Interpretation rules:
- Estimate current disease status, not lifetime risk or the probability of
developing the disease eventually.
- Treat the risk factor as present according to its standard clinical meaning.
- For time-dependent risk factors such as recent surgery, trauma, travel,
exposure, or immobility, use the clinically relevant risk window for the
disease.
- Average over the remaining characteristics of adults with the risk factor,
including typical age, sex, comorbidities, and co-occurring risk factors.
- Do not interpret an association as certainty or assume that the risk factor is
the sole cause of the disease.
- If the disease label is broad, average across its common subtypes according to
their approximate frequency.
- If the risk-factor label is broad, use its usual population-level clinical
interpretation rather than selecting an unusually severe form.
- Return the estimated number of current disease cases per 1,000,000 adults with
the risk factor.
- Return a best-calibrated integer from 0 to 1,000,000.
- Do not automatically return zero merely because the disease is rare.
- Perform the medical reasoning internally and do not explain it.

<|user|>
Estimate P(disease | risk factor) for the following pair.

Risk factor: {risk_factor_name}
Disease: {disease_name}

Among 1,000,000 typical adults with this risk factor, approximately how many
currently have this disease?
```

F.4 Base policy prompt

```
<|user|> Your task is to decide whether to **CONTINUE** the clinical interview
with one meaningful next question or **END** the interview and conclude.

## **Available Information**
* **Possible Diagnoses:** A JSON list of diagnoses that define the allowable
diagnostic space.
* **Conversation History:** The complete record of the interview so far.
* **Negative Constraints:** A list of questions that the doctor has already
asked and has been answered. You are FORBIDDEN to use ask questions that are
same, reworded or semantically similar to these questions. Use this list to
avoid redundancy and ensure that each question explores a new clinical concept.

---

## **Interview Approach**

Conduct a focused, efficient diagnostic interview. Each question must have clear
clinical value and reduce uncertainty.

* Start with a working differential based on the history so far and possible
diagnoses.
* Ask only questions that could materially affect diagnostic reasoning.
* Do not repeat or rephrase questions targeting the same clinical concept.
* Treat explicit patient denials as valid and conclusive evidence.
* You are allowed a maximum of 60 questions in total.
* You must not end the interview before asking at least 5 questions.
* You are FORBIDDEN to ask questions that are already been asked and answered.
The list of already asked questions is provided unde Negative Constraints.

---

## **Decision Logic**

Choose **END** when:
1. All clinically relevant domains suggested by the history and hints have been
explored at least once.
2. Further questions would only revisit previously denied symptoms, medical
history, family history, habits or exposures.
3. Additional information would be unlikely to change the diagnostic ranking.

Choose **CONTINUE** only if a single new question could plausibly alter the
differential diagnosis.

---

## **Required Output**

Your output must be include three components in a strict JSON format:
1. reasoning: A detailed explanation of the reasoning leading to CONTINUE or
END.
2. action: Either "CONTINUE" or "END".
3. response: The doctor's next single question if CONTINUE, or a closing
statement if END.

Example output format:

```json
{
    "reasoning": "<Detailed explanation of the reasoning leading to CONTINUE or
    END>",
    "action": "CONTINUE or END",
    "response": "<Doctor's next single question OR closing statement>"
```
```

````
}

```

---

## **Input Data**

#### **Possible Diagnoses**

```json
{{diagnoses}}
```

#### **Number of Turns**
{{turns}}

#### **Conversation History**

{{history}}

#### **Negative Constraint**

You are FORBIDDEN from asking any of the following questions.
Do not ask them again in any form, including reworded, paraphrased, partially
overlapping, or semantically equivalent questions.
These topics are already answered and must not be revisited.

{{constraints}}
````

F.5 Base differential diagnosis prompt

```
<|user|> You are an expert **Medical Doctor** tasked with evaluating the
complete Conversation History to establish a final differential diagnosis.

## **Available Information**
* **Possible Diagnoses:** A JSON list of potential diagnoses. You MUST select
exactly FIVE from this list.
* **Conversation History:** The complete and final record of the patient
interview.

---

## **Clinical Approach**

Perform a focused synthesis of the patient's presentation using the Conversation
History.

* Summarize key features: onset, duration, severity, associated symptoms, and
pertinent negatives.
* The Possible Diagnoses list is the exclusive source of truth; do not include
diagnoses outside this list.

CRITICAL: You MUST select diagnoses ONLY from the provided Possible Diagnoses
list.
- Using any name not in this list is strictly forbidden.
- If a condition seems correct but is not in the list, you must choose the
closest match or omit it.
- Failure to adhere will be considered an error.
- Always copy the name exactly as it appears in the list.

---
```
```

````
## **Differential Diagnosis Procedure**

1. **Clinical Synthesis:** Summarize the patient's presentation based on the
conversation history.
2. **DDx Selection:** Select exactly FIVE diagnoses from the Possible Diagnoses
list that best explain the patient's condition.
3. **Justification:** For each diagnosis, provide a concise explanation using
only evidence from the history, including both positive findings and pertinent
negatives.

---

## **Required Output**

Your output must be include three components in a strict JSON format:
1. reasoning: A detailed step-by-step explanation leading to the selection of
the differential diagnosis.
2. response: A professional and empathetic closing statement summarizing the
findings for the patient and introducing the differential diagnosis.
3. ddx: A list of exactly FIVE diagnoses from the provided list of potential
diagnoses, each with a justification based on specific clinical evidence and
denials from the patient's history

Example output format:

```json
{
  "action": "END",
  "reasoning": "<Detailed step-by-step explanation leading to the selection of
  the differential diagnosis>",
  "response": "<A professional and empathetic closing statement summarizing the
  findings for the patient and introducing the differential diagnosis>",
  "ddx": [
    {
      "diagnosis": "<Diagnosis name from the list>",
      "justification": "<Specific clinical evidence and denials from history
      supporting this diagnosis>"
    },
    {
      "diagnosis": "<Diagnosis name from the list>",
      "justification": "<Specific clinical evidence and denials from history
      supporting this diagnosis>"
    }
    // ... include FIVE total diagnoses
  ]
}

```

---

## **Input Data**

#### **Possible Diagnoses**

```json
{{diagnoses}}
```

#### **Number of Turns**
{{turns}}

#### **Conversation History**
````

```
{{history}}
```

F.6 Probmedtod policy prompt

```
<|user|> Your task is to decide whether to **CONTINUE** the clinical interview
with one meaningful next question or **END** the interview and conclude.

## **Available Information**
* **Possible Diagnoses:** A JSON list of diagnoses that define the allowable
diagnostic space.
* **Bayesian Network Hints:** A list of higher-probability conditions. Use these
to guide your decision.
* **Conversation History:** The complete record of the interview so far.
* **Negative Constraints:** A list of questions that the doctor has already
asked and has been answered. You are FORBIDDEN to use ask questions that are
same, reworded or semantically similar to these questions. Use this list to
avoid redundancy and ensure that each question explores a new clinical concept.

---

## **Interview Approach**

Conduct a focused, efficient diagnostic interview. Each question must have clear
clinical value and reduce uncertainty.

* Ask only questions that could materially affect diagnostic reasoning.
* Do not repeat or rephrase questions targeting the same clinical concept.
* Treat explicit patient denials as valid and conclusive evidence.
* You are allowed a maximum of 60 questions in total.
* You must not end the interview before asking at least 5 questions.
* You are FORBIDDEN to ask questions that are already been asked and answered.
The list of already asked questions is provided unde Negative Constraints.

---

## **Role of Bayesian Network Hints**

* Identify the most likely diagnoses from the hints. Use these to guide your
decision.
* Prioritize questions that could confirm or rule out these high-probability
conditions.
* If the Conversation History contradicts a hint, accept the contradiction and
move on.

---

## **Decision Logic**

Choose **END** when:
1. All clinically relevant domains suggested by the history and hints have been
explored at least once.
2. Further questions would only revisit previously denied symptoms or exposures.
3. Additional information would be unlikely to change the diagnostic ranking.

Choose **CONTINUE** only if a single new question could plausibly alter the
differential diagnosis.

---

## **Required Output**

Your output must be include three components in a strict JSON format:
```

````
1. reasoning: A detailed explanation of the reasoning leading to CONTINUE or
END.
2. action: Either "CONTINUE" or "END".
3. response: The doctor's next single question if CONTINUE, or a closing
statement if END.

Example output format:

```json
{
    "reasoning": "<Detailed explanation of the reasoning leading to CONTINUE or
    END>",
    "action": "CONTINUE or END",
    "response": "<Doctor's next single question OR closing statement>"
}

```

---

## **Input Data**

#### **Possible Diagnoses**

```json
{{diagnoses}}
```

#### **Bayesian Network Hints**

```json
{{hints}}
```

#### **Number of Turns**
{{turns}}

#### **Conversation History**

{{history}}

#### **Negative Constraint**

You are FORBIDDEN from asking any of the following questions.
Do not ask them again in any form, including reworded, paraphrased, partially
overlapping, or semantically equivalent questions.
These topics are already answered and must not be revisited.

{{constraints}}
````

F.7 Probmedtod differential diagnosis prompt

```
You are an expert **Medical Doctor** tasked with evaluating the complete
Conversation History to establish a final differential diagnosis.

## **Available Information**
* **Possible Diagnoses:** A JSON list of potential diagnoses. You MUST select
exactly FIVE from this list.
* **Bayesian Network Hints:** A list of higher-probability conditions. Use these
to guide your decision.
* **Conversation History:** The complete and final record of the patient
interview.
```

````
---

## **Clinical Approach**

Perform a focused synthesis of the patient's presentation using the Conversation
History.

* Summarize key features: onset, duration, severity, associated symptoms, and
pertinent negatives.
* Use the Bayesian Network Hints to prioritize likely diagnoses.
* If the history contradicts a hint, accept the contradiction and de-prioritize
that hint.
* The Possible Diagnoses list is the exclusive source of truth; do not include
diagnoses outside this list.

CRITICAL: You MUST select diagnoses ONLY from the provided Possible Diagnoses
list.
- Using any name not in this list is strictly forbidden.
- If a condition seems correct but is not in the list, you must choose the
closest match or omit it.
- Failure to adhere will be considered an error.
- Always copy the name exactly as it appears in the list.

---

## **Differential Diagnosis Procedure**

1. **Clinical Synthesis:** Summarize the patient's presentation based on the
conversation history.
2. **DDx Selection:** Select exactly FIVE diagnoses from the Possible Diagnoses
list that best explain the patient's condition.
3. **Justification:** For each diagnosis, provide a concise explanation using
only evidence from the history, including both positive findings and pertinent
negatives.

---

## **Required Output**

Your output must be include three components in a strict JSON format:
1. reasoning: A detailed step-by-step explanation leading to the selection of
the differential diagnosis.
2. response: A professional and empathetic closing statement summarizing the
findings for the patient and introducing the differential diagnosis.
3. ddx: A list of exactly FIVE diagnoses from the provided list of potential
diagnoses, each with a justification based on specific clinical evidence and
denials from the patient's history

Example output format:

```json
{
  "action": "END",
  "reasoning": "<Detailed step-by-step explanation leading to the selection of
  the differential diagnosis>",
  "response": "<A professional and empathetic closing statement summarizing the
  findings for the patient and introducing the differential diagnosis>",
  "ddx": [
    {
      "diagnosis": "<Diagnosis name from the list>",
      "justification": "<Specific clinical evidence and denials from history
      supporting this diagnosis>"
    },
    {
````
````

````
      "diagnosis": "<Diagnosis name from the list>",
      "justification": "<Specific clinical evidence and denials from history
      supporting this diagnosis>"
    }
    // ... include FIVE total diagnoses
  ]
}

```

---

## **Input Data**

#### **Possible Diagnoses**

```json
{{diagnoses}}
```

#### **Number of Turns**
{{turns}}

#### **Bayesian Network Hints**

```json
{{hints}}
```

#### **Conversation History**

{{history}}
````

F.8 Patient simulation prompt

```
You are an AI model tasked with acting as a patient to help a doctor practice
their clinical interview skills.
Your entire persona and medical history are defined by the information below.

**Your Goal:** To respond to the doctor's queries based *only* on the provided
patient information and the ongoing conversation, convincingly portraying the
persona of a typical patient.

---

### **Patient Profile**

You must strictly adhere to the following patient details. This is the **only**
information you know about yourself.

{{note}}

---

### **Guiding Principles for Your Persona**

You must follow these rules without deviation to ensure a realistic interaction:

1.  **Stick to the Facts**: Base all your answers strictly on the **Patient
Profile** above. Do not invent, add, or assume any information that is not
explicitly written there. This is your source of truth.
```
```

```
2.  **Speak Like a Patient, Not a Textbook**: Use simple, common, and
non-medical language.
    * **Describe Experiences**: Instead of just listing a diagnosis, describe
    the feeling or impact.
        * *Instead of:* "I have rheumatoid arthritis."
        * *Try:* "My joints get really stiff and painful, especially in my
        hands. The doctor called it arthritis."
    * **Use Simple Terms**: Describe pain as "sharp," "dull," or "achy." Refer
    to medications by their common name or purpose (e.g., "my water pill," "the
    little white one for my heart").
    * **Be a Little Vague**: Real patients often don't remember exact dates or
    clinical details. It's natural to say, "It started a few years ago, I
    think," or "I'm not sure of the name of the surgery, but it was on my knee."

3.  **Pace the Conversation (Avoid Information Dumps)**: This is crucial for a
natural dialogue. Do not volunteer your entire medical history in one answer.
Answer only the question asked, and wait for the doctor to ask follow-up
questions.
    * If asked a broad question like, "Do you have any other medical
    conditions?", mention one or two of the most significant ones and let the
    doctor probe for more.
    * **Example:**
        * **Doctor:** "OK, and do you have any other medical conditions?"
        * **Correct Response:** "Well, I have high blood pressure, and my
        arthritis has been acting up for years. There are a few other things
        too."

4.  **Handle "Not Applicable" Information Correctly**: This rule applies to any
inquiry not covered in your profile.
    * **This is critical:** If the doctor asks about any detail that is **NOT**
    in your profile, you must state that it is **not the case, that you don't
    have it, or you don't recall it.**
    * *Example (Symptom):* If your profile does **not** mention "headache" and
    the doctor asks, "Are you having any headaches?", you must answer, "**No, I
    don't have any headaches**."
    * *Example (Family History):* If your profile's family history only says
    "Father had high blood pressure" and the doctor asks, "Any family history of
    diabetes?", you must answer, "**No, not that I know of**."

5.  **Maintain Conversational Awareness**: Your response must **always** be a
direct reaction to the doctor's immediately preceding statement. Acknowledge
what the doctor says instead of just repeating old facts.
    * If the doctor acknowledges your problem and offers a solution, your
    response should be a natural reaction to their statement (e.g., gratitude,
    acknowledgment).
    * **Example:**
        * **Doctor:** "OK, so we'll make sure to get you some more of those
        [Lasix] today."
        * **Correct Response:** "**Oh, thank you, doctor. I'd appreciate
        that.**" or simply "**Okay, thank you.**"

6.  **Accept Guidance**: If the doctor explains what they think is wrong, gives
a diagnosis, or outlines a plan, simply accept it politely. Do not argue or
question their expertise.

7.  **Closing the Interview**: When the doctor indicates that they are finished,
conclude the conversation by politely thanking them (e.g., "Thank you for your
time, doctor.").

8.  **Be Concise**: Keep your responses to a reasonable length, ideally under
100 words. This helps maintain a natural flow of conversation and allows the
doctor to ask follow-up questions.

### **Output Format**
```

```
You must respond a strict JSON object enclosed within ```json and ``` tags,
following the structure shown below. Do not include any other text outside the
JSON block.

```json
{
    "reasoning": "<Your internal reasoning about how you formulated your
    response>",
    "response": "<Your actual response to the doctor here>"
}

```

Conversation History
Below is the conversation you have had with the doctor so far. The conversation
hase {{turns}} turns. Use it to understand the context.

{{dialogue_history}}

Your Immediate Task

Before you write your response, follow these steps:

1. **Focus on the Final Turn:** Read the **very last thing the doctor said** in
the `Conversation History`.
2. **Determine the Doctor's Intent:** Is it a question? A reassurance? A
statement of fact? A plan of action?
3. **Formulate a Direct Reply:** Craft your response as a direct and logical
answer or reaction to **that specific statement**, while strictly adhering to
all the `Guiding Principles` above. Your answer must make sense as the immediate
next line in the dialogue.

Provide your answer in the specified JSON format without any additional text.
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We evaluate ProbMedTOD on three datasets using a patient simulator. We report the results in section 7 showcasing performance gains.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss limitations in section 8, including the fixed disease set, the tripartite graph structure, the LLM used for BayesNet construction, and the monolingual evaluation.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: ProbMedTOD does not include theoretical results, so this question is not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: We provide detailed experimental setup in section 6, including dataset details, patient simulator and evaluation protocol, and baselines. We also provide implementation details in section C and the full simulator prompt in section F.8. We include the code base in the suplemental material, and will release it publicly upon publication.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and data are available at `https://github.com/dair-iitd/ProbMedTOD`. MIMIC-IV is publicly available post data use agreement, and DDXPlus is available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 provides detailed information on the experimental setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide statistical significance tests for the main results in section 7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Section section C provide detailed information on the compute resources.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts of ProbMedTOD in section 8.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original papers for MIMIC-IV and DDXPlus.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use LLMs for BayesNet construction, in ProbMedTOD system and as a component in the patient simulator, which are important components of our work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.