BridgeGap: Belief-Grounded Training for Medical Dialogue Agents

Anonymous ACL submission

Abstract

Although LLMs perform well on static medical QA, adapting them to multi-turn diagnostic dialogue remains challenging: agents forget confirmed symptoms, ask redundant questions, and fail to narrow differentials, largely because they lack an explicit representation of their evolving clinical understanding. Our key insight is that a structured belief state—encoding accumulated evidence, ranked differential diagnoses with confidence, and prioritized next questions—can simultaneously serve as a supervision target, a basis for rule-based reward computation, and a grounding context for deciding what to ask next. We propose BridgeGap, a two-stage framework in which supervised fine-tuning on silver-annotated belief states teaches the agent to maintain structured clinical reasoning, while reinforcement learning with a belief-progress reward refines questioning strategy through dense, environment-verified per-turn feedback without LLM judges or learned reward models. On HealthBench, BridgeGap (8B) surpasses Doctor-R1 (8B) by 2.84 points and the $4\times$ larger Baichuan-M2-32B by 5.85 points; on AgentClinic, it achieves 44.6% diagnostic accuracy in 10.5 average turns, compared to 37.6% in 15.5 turns for Doctor-R1. Across evaluations, BridgeGap advances interactive diagnosis while preserving factual knowledge on static medical benchmarks.

1 Introduction

Clinical diagnosis is a sequential reasoning task: physicians maintain a mental model of confirmed findings, open uncertainties, and the next most informative question, updating it with every patient response (Hauskrecht and Fraser, 2000). Large language models excel at static medical QA and exam-style benchmarks (Singhal et al., 2023; Nori et al., 2023), yet in interactive settings they forget confirmed symptoms, ask redundant questions, and fail to converge within a reasonable turn budget (Feng et al., 2025; Lai et al., 2026; Verma et al., 2026). We argue these failures share a common root: medical dialogue agents lack an *explicit, structured representation* of their evolving clinical understanding (Figure 1).

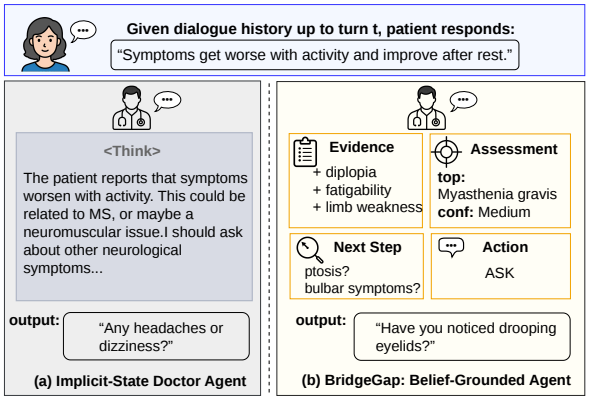

Figure 1: **(a)** Implicit reasoning conflates confirmed and uncertain findings, producing redundant questions. **(b)** BridgeGap grounds each action in an explicit belief state that tracks evidence, assessment, and next steps.

While belief states are central to task-oriented dialogue (Young et al., 2013; Mrkšić et al., 2017; Balaraman et al., 2021) and process supervision has been shown to outperform outcome-only signals in multi-step reasoning (Uesato et al., 2022; Lightman et al., 2024), current LLM-based medical dialogue agents keep clinical state implicit in free-form reasoning, opaque to turn-level verification. This produces three coupled failures: **(1) Information management.** Without an explicit evidence record, the agent cannot distinguish what has been established from what remains unknown, producing redundant questions that exhaust the turn budget. **(2) Strategic planning.** Without a structured assessment of diagnostic uncertainty, the agent has no principled basis for deciding *what to ask next* to narrow the differential. **(3) Training signal.** A dense per-turn reward must measure whether each question genuinely advanced the diagnostic pro-

cess, which itself requires the structured state the agent does not maintain.

Recent work attacks these problems from different angles, but existing reward designs typically trade off among *density*, *cost*, and *grounding in clinical state*, rarely providing all three at once. Doctor-R1 (Lai et al., 2026) obtains dense per-turn signals via an LLM judge, at the cost of expensive, version-sensitive scoring not anchored to a structured belief. IGFT (Verma et al., 2026) measures entity-level information gain—dense and cheap but blind to how new facts shift the differential. DoctorAgent-RL (Feng et al., 2025) uses rule-based F1 on final diagnoses, which is grounded but too sparse to shape intermediate questioning. All three operate over unstructured reasoning, leaving the reward signal only indirectly tied to diagnostic belief revision.

We introduce **BridgeGap** (Figure 1b): at every turn the agent generates a *structured clinical belief state* that bridges the gap between implicit internal reasoning and a clinical state that can be checked against the dialogue environment. The belief encodes accumulated evidence (confirmed, denied, pending), a ranked differential with confidence, and prioritized next questions. Generated *inside* the LLM rather than via an external state tracker, this representation plays three otherwise separately-engineered roles. In Stage 1 SFT, automatically generated (silver) belief annotations teach the agent to organize clinical information systematically. In Stage 2 RL, consecutive belief states drive a *belief-progress reward*—a rule-based per-turn signal verified against environment-extracted findings, requiring no LLM judge or learned reward model. At inference, the same belief grounds each response in explicit clinical context.

We evaluate across interactive and static benchmarks. On HealthBench (Arora et al., 2025), BridgeGap achieves the highest score among all evaluated open-source models including those up to 70B, outperforming Doctor-R1 by 2.84 points and UltraMedical-70B by over 12 points, with particularly large gains on Context Seeking (+7.13), a dimension closely aligned with systematic information tracking. On AgentClinic (Schmidgall et al., 2026), it reaches 44.6% diagnostic accuracy in 10.5 turns on average, compared to 37.6% in 15.5 turns for Doctor-R1, the strongest baseline evaluated. Static medical benchmarks (MedQA, MedMCQA, PubMedQA, MMLU) show no substantial degradation relative to the base model.

Our contributions are threefold:

1. We formalize a structured clinical belief state that simultaneously serves as supervision target, policy context, and reward substrate—to our knowledge, the first medical dialogue framework to unify these three roles in a single representation.
2. We design a belief-progress reward that is rule-based and verified against environment-extracted findings, providing a dense per-turn signal without learned or judge-based reward models.
3. Through extensive experiments and controlled ablations, we show that explicit belief structure and dense belief-progress reward provide *complementary* gains across response quality, factual knowledge retention, and interactive clinical diagnosis.

2 Related Work

Reasoning representations for medical dialogue. Clinical dialogue requires tracking established evidence, remaining uncertainty, and the next informative action, yet existing representations sacrifice either *structural parsability* or *end-to-end trainability*. Implicit free-text reasoning via chain-of-thought or hidden Think blocks (DeepSeek-AI, 2025; Chen et al., 2025; Lai et al., 2026) produces rich traces that resist turn-level reward computation and cross-turn consistency checks. Structured alternatives—dialogue state tracking with slot ontologies (Balaraman et al., 2021; Liu and Lane, 2017), external Bayesian belief engines (Kesmen et al., 2026), and rule-based clinical extraction pipelines (Pan et al., 2026)—are verifiable but tied to fixed schemas or modules separated from policy learning. Medical foundation models (Singhal et al., 2023; Baichuan-M2 Team, 2025; Chen et al., 2025) mainly improve knowledge and instruction following, while we make the agent's evolving clinical reasoning state explicit, trainable, and verifiable. BridgeGap generates a structured belief state *inside* the LLM, unifying supervision target, policy context, and reward substrate in one representation.

Per-step reward for medical dialogue RL. Process reward models have shown advantages over outcome-only signals in mathematical reasoning (Lightman et al., 2024). For medical dialogue, however, per-turn rewards must balance cost, reproducibility, and semantic granularity, and multi-turn online RL with per-step credit assignment remains rare in medical LLM-agent settings. LLM-

judge scoring (Lai et al., 2026) is rich but expensive and version-sensitive; entity-level information gain (Verma et al., 2026) and rule-based diagnosis F1 (Feng et al., 2025) are cheap but capture only fact acquisition or terminal correctness, not whether findings advance diagnostic reasoning. Rubric-based training (Wang et al., 2025) and RLVR (Su et al., 2025) pursue verifiable rewards, but the former requires task-specific rubric design while the latter primarily targets single-step or short-horizon settings. Our belief-progress reward instead compares consecutive belief states with fixed rules verified against environment-extracted findings, operating at the diagnostic reasoning level without a learned reward model.

Structured intermediate states as LLM training targets. Intermediate reasoning signals are increasingly used as *training targets*—from chain-of-thought traces supervised as auxiliary outputs (DeepSeek-AI, 2025) to process annotations driving per-step reward models (Uesato et al., 2022; Lightman et al., 2024). Turning such signals into structured states for clinical dialogue requires turn-level supervision that is hard to obtain at scale. Recent distillation (Wu et al., 2025) and LLM-as-annotator pipelines (Tan et al., 2024) can reduce annotation cost, but they often produce free-form traces that are unsuitable as targets for structured reward or consistency loss. Our silver pipeline extracts evidence labels via biomedical NER and prompts a teacher LLM only for assessment and next-step fields under an evidence-constrained JSON schema, producing structured belief labels with confidence-weighted filtering.

3 Method

3.1 Problem Formulation

We model multi-turn clinical consultation as a POMDP: the patient's true state s (diagnosis, full symptom profile) is hidden; at each turn t the agent observes the dialogue history h_t, chooses $a_t \in \{\text{ASK}, \text{DIAGNOSE}\}$, and aims to diagnose correctly within $T_{\max}$ turns. Under partial observability, optimal action selection requires maintaining a belief state—a sufficient statistic of past observations. Rather than leaving this belief implicit in the LLM's context window, we make it explicit and structured (§3.2), enabling it to simultaneously serve as supervision target, reward substrate, and policy context (Figure 2).

3.2 Structured Belief State

At each turn t, the agent produces a structured belief state b_t summarizing its current clinical understanding. The schema mirrors the minimum state required for sequential diagnosis: *evidence* tracks what has been established, *assessment* tracks the current diagnostic hypothesis and uncertainty, *next step* specifies what information would best reduce that uncertainty, and *action* makes the stopping decision explicit. The belief consists of three information groups plus a control field:

Evidence. Three sets tracking what the agent knows: `confirmed` (findings affirmed by the patient), `denied` (explicitly negated), and `pending` (mentioned but not yet available, e.g., lab results). Evidence accumulates incrementally across turns.

Assessment. The agent's diagnostic reasoning: a `top` diagnosis (single most likely), 2–4 `alternatives`, and a discrete `confidence` level (LOW/MEDIUM/HIGH). We use qualitative confidence rather than calibrated probabilities, which are unrealistic for generative LMs and would create self-referential reward issues.

Next Step. Prioritized questions for the next turn: 1–3 `critical` questions that would most effectively distinguish `top` from `alternatives`, and 1–3 `helpful` supporting questions.

Action. A binary control field (ASK or DIAGNOSE): DIAGNOSE indicates the agent judges sufficient evidence has been gathered; ASK indicates further inquiry is needed, with targets drawn from `next_step`.

Output format. At each turn, the model generates two blocks sequentially:

1. A hidden `<PLAN>` block containing the structured belief state (evidence, assessment, next step) and the action decision.
2. A visible `<UTTERANCE>` block containing the actual response to the patient.

Only the `<UTTERANCE>` is transmitted to the environment. The `<PLAN>` serves as the agent's internal reasoning workspace — used for supervision, reward computation, and episode termination detection, but never exposed to the patient. When `action`=DIAGNOSE, the environment terminates the episode and evaluates the diagnosis in `assessment.top` against ground truth.

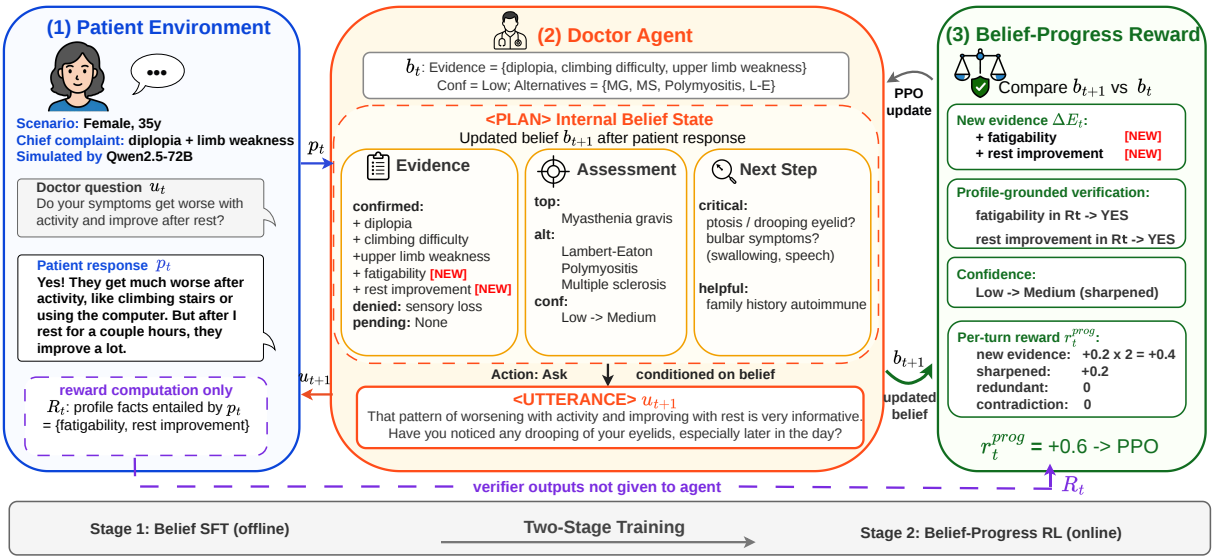

Figure 2: One interaction turn during Stage 2. The patient discloses new findings (left); the agent updates its structured belief b_{t+1} (center) and generates a belief-conditioned utterance. The reward module (right) compares consecutive beliefs against profile-grounded fact verification, producing a dense per-turn signal for PPO. In Stage 1, the same belief structure serves as offline SFT supervision; verifier outputs are never exposed to the agent.

3.3 Silver Belief Annotation

Training the agent to produce structured belief states requires turn-level supervision. Since expert annotation at scale is prohibitively expensive, we construct 35k silver-labeled samples through a two-step pipeline applied to three public dialogue sources (details in Appendix A).

Data sources. We draw from MediTOD (Saley et al., 2024) (231 multi-turn consultations with gold slot annotations), MTS-Dialog (Ben Abacha et al., 2023) (filtered history-taking segments), and MedDialog-EN (Zeng et al., 2020) (226k single-turn consultations from which we extract patient profiles and instantiate simulated multi-turn consultations). None of these sources overlaps with our evaluation benchmarks.

Step 1: Evidence annotation. We apply MedSpaCy (Eyre et al., 2021) (biomedical NER + UMLS concept linking) with the **ConText algorithm** (Harkema et al., 2009) for assertion classification. Each extracted clinical entity is categorized as `confirmed` (affirmed by patient), `denied` (negated), or `pending` (hypothetical or doctor-proposed). On MediTOD's gold annotations, this pipeline achieves 87% precision and 82% recall (Appendix A.2).

Step 2: Assessment and Next Step. Given the extracted evidence and dialogue prefix, we prompt GPT-4o (`gpt-4o-2024-11-20`) with structured output constraints (JSON Schema) to generate the `assessment` and `next_step` fields. The LLM receives evidence as input but *not* the ground-truth diagnosis, ensuring labels reflect reasoning from available information only. GPT-4o produces structured intermediate belief labels under an evidence-constrained JSON schema, rather than free-form responses or final-diagnosis targets.

Quality filtering. Each silver label carries a confidence score that weights the SFT loss, so low-confidence labels contribute less. A stratified human audit by two clinically-trained annotators confirms that label accuracy increases monotonically with the confidence score; high-confidence labels reach 96% evidence accuracy and 90% assessment accuracy (Appendix A.5).

3.4 Two-Stage Training

3.4.1 Stage 1: Belief SFT

Starting from Qwen3-8B-Instruct, we fine-tune on 35k silver-annotated dialogue trajectories. Each sample consists of a dialogue prefix paired with the target: `<PLAN>` (silver belief) followed by `<UTTERANCE>` (ground-truth next doctor utterance).

Objective.

$$\mathcal{L}_{\text{S1}} = \mathcal{L}_{\text{SFT}} + \lambda_B \cdot \mathcal{L}_{\text{belief}} \quad (1)$$

where $\mathcal{L}_{\text{SFT}}$ is standard next-token cross-entropy over the full `<PLAN>`+`<UTTERANCE>` sequence,

and λ_B=0.3 weights the auxiliary consistency term.

Belief consistency loss. $\mathcal{L}_{\text{belief}}$ prevents unjustified retraction of established findings: for each multi-turn sample at turn t, any finding present in `confirmed` at $t-1$ receives additional cross-entropy loss to ensure it remains confirmed at t. Patient-initiated contradictions are exempted via the contradiction detection mechanism described above.

Policy conditioning. The model generates <UTTERANCE> *after* <PLAN>, so each action is conditioned on the predicted belief: $a_t \sim \pi(\cdot \mid h_t, b_t)$ where b_t denotes the structured content of <PLAN>. The belief thus serves as a structured reasoning context that organizes clinical thinking before action selection. This distinguishes our approach from auxiliary prediction tasks where the intermediate output is not used for decision-making.

3.4.2 Stage 2: Belief-Progress RL

Stage 2 refines the agent's questioning strategy through reinforcement learning in an online multi-turn environment. At the start of each episode, we sample one structured clinical profile from 2,400 profiles constructed from MedDialog-EN (Zeng et al., 2020); the same profile conditions the frozen Qwen2.5-72B Patient Agent and provides the finite fact set for reward verification. Each profile enumerates its clinical facts—positive findings, pertinent negatives, history, labs—with canonical names that the Patient Agent reveals naturally across turns. This online setup allows RL to optimize questioning behavior without pre-annotated belief labels. Profile construction and contamination filtering are detailed in Appendix C.1.

Reward: belief progress. The reward measures whether each action meaningfully advanced clinical reasoning, computed from rule-based comparison of consecutive belief states. All reward values are **fixed constants**; no learned reward model is required. The belief-progress reward at turn t is:

$$\begin{aligned} r_t^{\text{prog}} = \alpha \sum_{e \in \mathcal{E}_t^{\text{new}}} & \mathbb{1}[e \in \mathcal{R}_t] \\ & - \beta \cdot \mathbb{1}[\text{redund}_t] - \gamma \cdot \mathbb{1}[\text{contra}_t] \\ & + \delta \cdot \mathbb{1}[\text{sharp}_t] \end{aligned} \tag{2}$$

where $\mathcal{E}_t^{\text{new}}$ denotes findings newly added to the confirmed set between consecutive beliefs b_t and b_{t+1}, aligned to the canonical profile-fact space, and $\mathcal{R}_t$ is determined by **profile-grounded closed-set verification**: a frozen NLI model (Deka et al., 2023) (`PubMedBERT-MNLI-MedNLI`) checks which facts from the patient's finite profile $\mathcal{F}$ are entailed by p_t (Appendix C.2). Constants are α=0.2, β=0.1, γ=0.3, δ=0.2. The largest penalty goes to contradiction (γ=0.3) because unsupported belief reversals directly undermine state consistency.

Interaction timeline. The per-turn sequence is:

$$h_t \;\rightarrow\; (b_t,\, u_t) \;\rightarrow\; p_t \;\rightarrow\; (b_{t+1},\, \mathcal{R}_t) \;\rightarrow\; r_t^{\text{prog}}$$

At turn t, the agent observes dialogue history h_t and generates belief b_t together with utterance u_t. The patient responds with p_t, from which two quantities are derived independently: the agent generates updated belief b_{t+1} (using only p_t and history), while the environment-side verifier extracts $\mathcal{R}_t$ (facts entailed by p_t). The reward r_t^{prog} then compares b_t against b_{t+1} using $\mathcal{R}_t$ as ground truth—crucially, $\mathcal{R}_t$ is never exposed to the agent. Verification pipeline details are in Appendix C.2; a human audit confirms pipeline–human agreement at κ=0.82 (Appendix C.3).

Terminal reward. When `action`=DIAGNOSE:

$$r_T = \begin{cases} +1.0 & \texttt{top} \text{ matches ground-truth} \\ +0.5 & \text{ground-truth in } \texttt{alternatives} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

If the agent fails to select DIAGNOSE within $T_{\max}$ turns, the episode terminates with $r_T = -0.5$, discouraging indefinite information gathering.

4 Experiments

4.1 Setup

Benchmarks. We evaluate on three complementary benchmarks: (1) **HealthBench** (Arora et al., 2025), 5,000 multi-turn healthcare conversations scored across seven themes and five behavioral axes by physician-authored rubrics. (2) **Static medical benchmarks**: MedQA (US 4-option) (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and six MMLU medical subsets (Hendrycks et al., 2021), ensuring dialogue training preserves factual knowledge. (3) **AgentClinic** (Schmidgall et al., 2026), a multimodal agent benchmark where the doctor agent interacts with patient and measurement agents in a simulated clinical environment.

Table 1: Results on HealthBench Main. Best open-source results in **bold**. BridgeGap obtains the best open-source average score. Columns are grouped into Themes (scenario categories) and Axes (behavioral skills).

		Avg.	Theme							Axis				
Model	**Scale**	**Score**	**Emerg. Referrals**	**Health Data T.**	**Commu-nication**	**Global Health**	**Hedging**	**Context Seeking**	**Complex Resp.**	**Acc.**	**Comm. Quality**	**Instr. Foll.**	**Context Aware.**	**Comp-leteness**
Open-Source Models (≥32B)														
Baichuan-M2	32B	33.04	20.49	9.07	21.86	18.63	17.06	28.51	24.18	34.22	57.74	52.68	46.52	40.29
OpenBioLLM-70B	70B	18.78	33.91	12.62	23.12	9.44	16.23	9.51	30.58	24.09	48.35	38.05	37.54	20.11
HuatuoGPT-o1-70B	70B	21.04	36.92	18.02	24.18	11.79	20.03	12.63	29.42	26.08	55.21	49.72	38.23	22.91
Med42-v2-70B	70B	25.79	41.22	24.39	35.91	13.25	21.14	15.83	33.46	30.87	55.76	51.71	41.61	26.69
UltraMedical-70B	70B	26.57	38.31	23.48	31.18	19.68	17.86	16.71	**45.47**	32.34	50.89	40.96	45.72	27.36
Open-Source Models (7B–8B)														
HippoLlama	7B	2.44	0.61	0.39	1.82	0.72	4.16	0.42	12.34	6.89	9.76	20.41	15.88	5.03
HippoMistral	7B	6.08	9.82	2.24	5.94	3.51	5.83	3.17	13.49	11.29	23.58	18.05	22.12	9.48
BioMistral	7B	6.12	12.76	2.85	5.83	3.31	7.36	3.09	12.08	11.57	19.38	24.21	23.41	8.39
OpenBioLLM-8B	8B	8.34	18.05	6.58	6.12	4.96	7.61	2.69	16.21	13.17	25.43	24.57	23.49	10.92
Med42-v2	8B	15.11	28.17	18.77	13.86	7.61	14.03	9.08	32.21	20.91	45.39	42.72	30.84	18.04
DoctorAgent-RL	7B	15.58	33.76	14.82	17.84	7.52	15.51	8.04	30.27	20.06	50.32	38.28	32.01	18.31
HuatuoGPT-o1	8B	16.42	33.89	18.26	17.49	8.36	14.89	11.14	26.28	21.36	53.96	41.83	33.55	17.29
UltraMedical	8B	22.03	46.83	19.34	26.39	13.56	19.55	18.06	26.98	25.24	57.68	44.42	40.51	27.15
Doctor-R1	8B	36.05	54.21	28.83	46.97	24.57	33.13	26.22	34.17	37.71	63.96	53.84	49.03	40.51
BridgeGap	**8B**	**38.89**	**58.48**	**31.51**	**53.39**	**27.48**	**36.97**	**33.35**	35.03	**42.43**	**69.34**	**56.75**	**55.30**	**45.69**

Table 2: Static medical benchmark results. Best 8B-scale results in **bold**. MMLU reports six medical subsets.

					MMLU						
Model	**Scale**	**MedQA (US 4-opt)**	**MedMCQA (Dev)**	**PubMedQA (Reasoning)**	**Avg.**	**Clinical Knowledge**	**Medical Genetics**	**Anatomy**	**Professional Medicine**	**College Biology**	**College Medicine**
Open-Source Models (≥32B)											
OpenBioLLM-70B	70B	78.16	74.01	78.97	91.02	92.93	93.20	83.90	93.75	93.83	85.75
HuatuoGPT-o1-70B	70B	83.30	73.60	80.60	87.97	86.04	91.00	82.96	93.38	94.44	79.19
UltraMedical-70B	70B	82.20	72.30	78.80	88.03	86.40	91.00	82.20	92.30	89.60	86.70
Med42-v2-70B	70B	79.10	73.20	78.80	84.48	82.64	89.00	76.30	88.60	92.36	78.03
Baichuan-M2-32B	32B	79.47	66.12	64.00	87.05	84.53	92.00	81.48	91.54	92.36	80.92
Open-Source Models (7B–8B)											
Doctor-R1	8B	**83.50**	58.38	75.80	85.03	86.04	**88.00**	**76.30**	86.40	88.89	83.24
UltraMedical-8B	8B	73.30	**61.50**	77.00	77.50	78.90	78.00	74.10	83.80	78.50	71.70
HuatuoGPT-o1-8B	8B	72.60	60.40	**79.20**	78.05	77.36	86.00	71.85	80.88	81.94	71.68
Med42-v2-8B	8B	62.84	61.34	68.40	73.65	75.85	79.00	69.63	75.37	76.39	65.32
OpenBioLLM-8B	8B	58.99	56.91	74.12	77.08	76.10	86.10	69.83	78.21	84.21	68.04
DoctorAgent-RL	7B	58.52	56.06	77.40	75.76	76.60	80.00	75.56	74.63	84.72	66.47
Qwen3-8B-Instruct	8B	65.59	58.09	74.20	79.61	78.87	80.00	69.63	82.35	86.11	78.61
BridgeGap	**8B**	82.36	60.52	76.80	**85.59**	**86.42**	86.00	75.56	**87.87**	**90.28**	**84.39**

Baselines. We compare against open-source models spanning two scales: *≥32B*: Baichuan-M2-32B (Baichuan-M2 Team, 2025), UltraMedical-70B (Zhang et al., 2024), HuatuoGPT-o1-70B (Chen et al., 2025), Med42-v2-70B (Christophe et al., 2024), OpenBioLLM-70B (Pal, 2024). *7B–8B*: Doctor-R1 (Lai et al., 2026), DoctorAgent-RL (Feng et al., 2025), HuatuoGPT-o1-8B (Chen et al., 2025), UltraMedical-8B (Zhang et al., 2024), Med42-v2-8B (Christophe et al., 2024), OpenBioLLM-8B (Pal, 2024), BioMistral (Labrak et al., 2024), HippoLlama (Acikgoz et al., 2024), HippoMistral (Acikgoz et al., 2024). We also report Qwen3-8B-Instruct (our base model) on static benchmarks for reference. Evaluation prompts and parameters are detailed in Appendix D.

Implementation. BridgeGap fine-tunes Qwen3-8B-Instruct in two stages: SFT on 35k trajectories with GPT-4o-generated belief annotations, followed by PPO against a frozen Qwen2.5-72B patient simulator (Appendix C).

4.2 Overall Performance

HealthBench. Table 1 presents results on HealthBench Main. BridgeGap achieves an average score of 38.89, surpassing Doctor-R1 (36.05) by 2.84 points and the 4× larger Baichuan-M2-32B (33.04) by 5.85 points. The improvement is uneven across dimensions, reflecting the method's design: gains are largest on Context Seeking (+7.13 over Doctor-R1), Communication (+6.42), and Context Awareness (+6.27)—precisely the axes where belief-grounded planning contributes most—while knowledge-intensive themes such as Global Health (+2.91) and Complex Responses (+0.86) see modest gains limited by the 8B model's capac-

Table 3: AgentClinic-MedQA Extended results (213 cases). **Conf.**: patient agent's post-consultation confidence in the doctor's assessment (1–10 scale, higher is better). **Early≤5t**: percentage diagnosed correctly within 5 turns.

Model	Acc.	Turns↓	Conf.	Early≤5t
Doctor-R1	37.6	15.5	8.4	5.2
DoctorAgent-RL	17.1	15.1	6.0	1.9
HuatuoGPT-o1	16.3	18.36	8.27	0.9
Med42-v2	11.4	19.2	8.1	0.0
UltraMedical	0.5	19.9	3.3	0.0
BridgeGap	**44.6**	**10.5**	**8.8**	**10.3**

ity. BridgeGap leads all open-source models on six of seven Themes and on all five Axes.

Static medical benchmarks. Table 2 verifies that dialogue-focused training does not degrade factual medical knowledge. BridgeGap achieves an MMLU-Medical weighted average of 85.59, slightly exceeding Doctor-R1 (85.03) and approaching 32B-scale models (Baichuan-M2: 87.05). Gains concentrate on clinically-oriented subjects (Clinical Knowledge, Professional Medicine, College Medicine) where structured belief training provides transferable reasoning.

4.3 Analysis on AgentClinic

We conduct detailed analysis on AgentClinic-MedQA Extended, an interactive benchmark where the doctor agent must decide both *what to ask* and *when enough evidence has been gathered* to diagnose—directly testing the behavior targeted by BridgeGap.

Diagnostic accuracy and efficiency. Table 3 compares open-source medical agents under the same 20-turn interaction budget. BridgeGap reaches 44.6% diagnostic accuracy, improving over Doctor-R1 by 7.0 absolute points while using 5.0 fewer turns on average (10.5 vs. 15.5). The gain is not merely a final-answer effect: BridgeGap also doubles early diagnostic accuracy within five turns (10.3% vs. 5.2%), suggesting that the predicted `next_step.critical` field helps the agent ask discriminative questions earlier.

Diagnosis dynamics. Figure 3(a) shows that the improvement stems from a fundamentally different interaction pattern. Doctor-R1 accumulates correct diagnoses slowly and exhibits a large jump at the 20-turn limit (26.3% to 37.6%), indicating

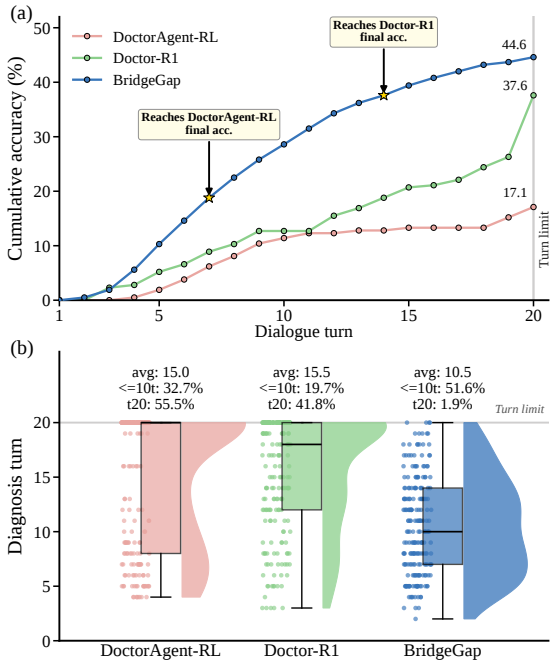

Figure 3: Diagnosis dynamics on AgentClinic-MedQA Extended. **(a)** Cumulative accuracy over turns. **(b)** Raincloud plot (jittered samples, box plot, half-violin) of diagnosis turn distribution.

heavy reliance on forced terminal guessing. In contrast, BridgeGap reaches 10.3% accuracy by turn 5, matches Doctor-R1's final accuracy by turn 14, and shows only a small turn-20 jump (43.7% to 44.6%). The turn distribution in Figure 3(b) further confirms this: BridgeGap diagnoses 51.6% of cases within 10 turns with only 1.9% reaching the turn limit, whereas Doctor-R1 exhausts 20 turns in 41.8% of cases. This confirms that the belief state improves not only final correctness but also stopping behavior: the agent recognizes when sufficient evidence has been collected rather than relying on end-of-dialogue guessing. A detailed case study comparing three models' dialogue strategies on the same patient is provided in Appendix E.

4.4 Ablation Study

We ablate key components on both HealthBench and AgentClinic to isolate each design choice's contribution. All variants use Qwen3-8B-Instruct as the base model: **ID 1** — dialogue SFT only (standard next-token loss on doctor utterances, no belief supervision); **ID 2** — adds structured belief state (Stage 1 with $\mathcal{L}_{\text{belief}}$); **ID 3** — adds interactive RL with terminal-only reward; **ID 4** — full BridgeGap with per-turn belief-progress reward.

Table 4: Ablation study on HealthBench and AgentClinic. Each row incrementally adds one component.

	Components				HealthBench					AgentClinic	
ID	Dialogue SFT	Belief Structure	Interactive RL	Belief-Prog. Reward	Avg	Acc.	Context Aware.	Comm. Quality	Comp-leteness	Acc.	Turns↓
1	✓	✗	✗	✗	23.41	26.18	35.72	52.83	25.14	15.5	18.7
2	✓	✓	✗	✗	34.26	37.51	50.18	64.27	39.83	32.4	16.2
3	✓	✓	✓	✗	35.94	39.12	51.47	65.83	41.26	36.2	13.8
4	✓	✓	✓	✓	**38.89**	**42.43**	**55.30**	**69.34**	**45.69**	**44.6**	**10.5**

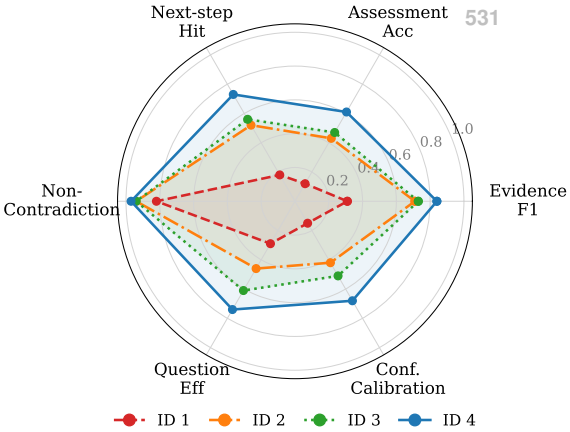

Figure 4: Belief quality metrics across ablation variants (Table 4) on AgentClinic episodes.

Belief structure is the core. The largest gain comes from adding the structured belief state (#1→#2): +10.85 HealthBench Avg with Context Awareness seeing the biggest jump (+14.46), and +16.9 AgentClinic accuracy. Notably, Completeness also improves by +14.69, indicating that belief-structured supervision teaches the model to cover more clinically relevant dimensions per response. This confirms that an explicit representation of evolving clinical knowledge—not merely more training data or RL—is the primary driver of both response quality and diagnostic success.

Dense reward unlocks RL. Terminal-only RL (#2→#3) adds only +1.68 HealthBench Avg, showing that sparse end-of-dialogue signals provide limited learning signal for response quality. However, it already reduces turns on AgentClinic by 2.4. Replacing terminal reward with per-turn belief-progress (#3→#4) yields +2.95 HealthBench Avg and +8.4 AgentClinic accuracy while further reducing turns by 3.3. The dense reward enables the agent to optimize each questioning step rather than relying solely on final-diagnosis feedback.

Belief quality analysis. Figure 4 reveals *how* each component improves belief fidelity, measured by six metrics (Evidence F1, Assessment Acc, Next-step Hit, Non-Contradiction, Question Efficiency, Conf. Calibration; definitions in Appendix B). (1) Adding belief structure (#1→#2) sharply improves Evidence F1 (+0.40) and Assessment Accuracy (+0.31), confirming that Stage 1 SFT teaches the model to identify and organize clinical facts—directly explaining the large end-to-end gains. (2) Terminal RL (#2→#3) primarily boosts Question Efficiency (+0.15) while leaving Evidence F1 and Assessment Acc nearly unchanged—terminal reward incentivizes *what to ask* but does not improve *how to track*. (3) Belief-progress reward (#3→#4) lifts Evidence F1 (+0.11), Next-step Hit (+0.17), and Confidence Calibration (+0.17) simultaneously, demonstrating that the dense belief-level reward directly improves belief fidelity, which in turn drives the end-to-end gains in Table 4.

5 Conclusion

We presented BridgeGap, a two-stage training framework that uses a structured clinical belief state as both SFT supervision target and RL reward signal for medical dialogue agents. The belief state—comprising confirmed evidence, diagnostic assessment, and prioritized next steps—provides a unified mechanism for state tracking, action selection, and dense reward computation. Our belief-progress reward is fully rule-based with environment-side verification, eliminating the need for LLM judges or learned reward models.

Experiments across interactive and static benchmarks demonstrate that BridgeGap outperforms all evaluated open-source baselines on interactive clinical tasks, with the largest gains on context-aware metrics that directly reflect belief tracking quality. Ablation studies confirm that the structured belief state is the primary contributor, and that dense belief-progress reward unlocks RL potential beyond sparse terminal signals alone.

Limitations

No empathy optimization. The reward targets diagnostic reasoning efficiency without explicitly optimizing for empathetic communication or patient emotional needs.

English-only scope. All benchmarks, training corpora, and profile facts are in English. Extending to multilingual clinical scenarios would require language-specific NLI models and culturally adapted patient profiles, which remain untested.

Simulated interaction. All interactive experiments rely on LLM patient simulators—a frozen Qwen2.5-72B agent during training, and benchmark-provided agents (separate from training) during evaluation—rather than real patients. Results should be interpreted as benchmark evidence; clinical validation with real patients remains future work.

References

Emre Can Acikgoz, Osman Batur İnce, Rayene Bench, Arda Anıl Boz, Ilker Kesen, Aykut Erdem, and Erkut Erdem. 2024. Hippocrates: An open-source framework for advancing large language models in healthcare. *arXiv preprint arXiv:2404.16621*.

Rahul K. Arora, Jason Wei, Rebecca Hicks Soskin, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Baichuan-M2 Team. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. 2021. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 239–251.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2291–2302.

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

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2023. Multiple evidence combination for fact-checking of health-related information. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks (BioNLP)*, pages 237–247. Association for Computational Linguistics.

Hannah Eyre, Alec B. Chapman, Kelly S. Peterson, Jianlin Shi, Patrick R. Alba, Makoto M. Jones, Tamara L. Box, Scott L. DuVall, and Olga V. Patterson. 2021. Launching into clinical space with medspacy: A new clinical text processing toolkit in python. *AMIA Annual Symposium Proceedings*, 2021:438–447.

Yichun Feng, Jiawei Wang, Lu Zhou, Yikai Zheng, Zhen Lei, and Yixue Li. 2025. Real-world doctor agent with proactive consultation through multi-agent reinforcement learning. *arXiv preprint arXiv:2505.19630*.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851.

Milos Hauskrecht and Hamish Fraser. 2000. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artificial Intelligence in Medicine*, 18(3):221–244.

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

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yusuf Kesmen, Fay Elhassan, Jiayi Ma, Julien Stalhandske, Yena Chang, David Sasu, Alexandra Kulinkina, Akhil Arora, Lars Klein, and Mary-Anne Hartley. 2026. Mobayes: A modular bayesian framework for separating reasoning from language in conversational clinical decision support. *arXiv preprint arXiv:2604.20022*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. 2026. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *International Conference on Learning Representations (ICLR)*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*.

Bing Liu and Ian Lane. 2017. An end-to-end trainable neural network model with belief tracking for task-oriented dialog. In *Interspeech 2017*, pages 2506–2510.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of ACL*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*.

Ankit Pal. 2024. Llama3-openbiollm-70b. Hugging Face model card.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pages 248–260. PMLR.

Zhenhai Pan, Yan Liu, and Jia You. 2026. Proactive knowledge inquiry in doctor-patient dialogue: Stateful extraction, belief updating, and path-aware action planning. *arXiv preprint arXiv:2603.17425*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of EMNLP*, pages 1586–1596.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam. 2024. 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 (EMNLP)*.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. 2026. Agentclinic: a multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*.

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

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding RL with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–957.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Tanvi Verma, Yang Zhou, Rick Siow Mong Goh, and Yong Liu. 2026. Aligning medical conversational ai through online reinforcement learning with information-theoretic rewards. *arXiv preprint arXiv:2601.17828*.

Pengkai Wang, Qi Zuo, Pengwei Liu, Zhijie Sang, Congkai Xie, and Hongxia Yang. 2025. Infimed-orbit: Aligning llms on open-ended complex tasks via rubric-based incremental training. *arXiv preprint arXiv:2510.15859*.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jin-Fang Hu, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical:

Building specialized generalists in biomedicine. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

A Silver Belief Annotation Pipeline

A.1 Data Sources and Scale

We construct 35k silver-annotated training samples from three sources:

1. **MediTOD** (Saley et al., 2024): 231 multi-turn consultations (avg. 48 doctor turns each), focused on respiratory conditions. Gold turn-level slot/intent/status annotations allow direct mapping to our belief structure and serve as NER pipeline validation. Multi-prefix sampling yields ∼5k samples.
2. **MTS-Dialog** (Ben Abacha et al., 2023): 1,201 section-level dialogue fragments from real clinical encounters. We retain only information-gathering sections (GENHX, CC, PASTMEDICALHX; ∼477 fragments, avg. 9.6 doctor turns). Paired clinical note text enables post-hoc assessment verification. Yields ∼2k samples.
3. **MedDialog-EN** (Zeng et al., 2020): 226k single-turn patient consultations spanning cardiology, dermatology, gastroenterology, neurology, and more. Used in two ways: (a) 10k richly-described patient messages become turn-1 belief training samples; (b) 4k case profiles extracted from patient messages drive synthetic multi-turn dialogue generation (GPT-4o as doctor, Qwen2.5-72B as patient), yielding ∼18k samples.

Contamination control. None of these sources overlap with our evaluation benchmarks. We additionally run embedding-based deduplication (cosine similarity < 0.8) against MedQA and AgentClinic case pools.

A.2 Step 1: Evidence Annotation

We use MedSpaCy with three components: (1) biomedical NER (`en_core_sci_lg`), (2) UMLS concept linking via QuickUMLS (threshold $= 0.7$), and (3) the ConText algorithm (Harkema et al., 2009) for assertion classification.

Each entity is classified as:

- `confirmed`: ConText = AFFIRMED and speaker = patient
- `denied`: ConText = NEGATED (any speaker)
- `pending`: ConText = HYPOTHETICAL, or entity appears in a doctor question/suggestion

Evidence accumulates incrementally: $\text{evidence}_t = \text{evidence}_{t-1} \cup \text{new_entities}_t$. Status transitions follow monotonicity constraints (pending → confirmed is allowed; confirmed → denied triggers contradiction detection).

Validation. On MediTOD's gold annotations, our NER pipeline achieves 87% precision and 82% recall for entity extraction + assertion classification.

A.3 Step 2: LLM-Generated Assessment

Given the dialogue prefix and extracted evidence, GPT-4o generates the `assessment`, `next_step`, and `utterance` fields via OpenAI Structured Outputs (JSON Schema enforcement, guaranteeing 100% format compliance).

Key design choices:

- The LLM receives evidence as structured input, reducing hallucination of unsupported findings.
- Ground-truth diagnosis is *never* provided in the prompt.
- For multi-turn sources (MediTOD, MTS-Dialog, synthetic), the `utterance` target is the actual next doctor turn from the dialogue (not LLM-generated).

A.4 Quality Filtering and Confidence Scoring

Each silver label receives a confidence score:

$$\text{score} = 0.4 \cdot \text{evidence_conf} + 0.6 \cdot \text{assessment_conf}$$

where evidence confidence is the mean UMLS similarity of extracted entities multiplied by the UMLS linkage coverage ratio, and assessment confidence maps the LLM's self-reported confidence (Low=0.3, Medium=0.6, High=0.9) with a consistency penalty if the top diagnosis contradicts the evidence.

Samples are discarded if: (1) no entities extracted, (2) LLM refuses to diagnose, (3) critical questions repeat confirmed evidence, or (4) score < 0.2. The SFT loss weight is $w = \min(1, \text{score}/0.7)$.

A.5 Silver Label Quality Audit

To validate silver label quality, we stratified 1,000 samples by confidence score into four groups (250 each: high $\geq$0.78, medium $[0.55, 0.78)$, low $[0.20, 0.55)$, and discarded $<$0.20) and conducted a two-tier audit. Two clinically-trained annotators (an internal medicine resident and an emergency medicine attending) independently reviewed 200 core samples (50 per stratum) across three dimensions: evidence status accuracy, assessment plausibility, and next-step usefulness. Disagreements

were adjudicated by a senior internist. An additional 800 samples were evaluated via automated heuristics with 50-sample human cross-validation.

Table 5: Silver label accuracy by confidence stratum (200 core samples, post-adjudication).

Stratum	Evid.	Assess.	Next-step
High (≥0.78)	96.0%	90.0%	86.0%
Medium [0.55, 0.78)	92.0%	82.0%	78.0%
Low [0.20, 0.55)	84.0%	72.0%	68.0%
Discarded (<0.20)	68.0%	48.0%	44.0%
Retained (avg.)	**90.8%**	**81.3%**	**77.3%**

Table 6: Inter-annotator agreement (Cohen's κ, pre-adjudication, n=200).

Stratum	Evid. κ	Assess. κ	Next-step κ
High	0.847	0.683	0.715
Medium	0.762	0.610	0.652
Low	0.715	0.561	0.604
Discarded	0.683	0.485	0.541
Overall	**0.741**	**0.596**	**0.640**

Three patterns emerge: (1) accuracy is strictly monotonic with confidence, validating the score as a quality proxy; (2) discarded samples are substantially worse than retained ones (evidence 68% vs. 84%, a 16pp gap), confirming that the filtering threshold removes genuinely low-quality labels; and (3) the difficulty gradient Evidence > Assessment > Next-step aligns with the increasing subjectivity of each dimension. On the 800-sample automated tier, heuristic pass rates track the core-layer accuracy within ±2pp across all strata, with a 50-sample cross-validation showing 86% agreement between heuristic and human judgments on evidence status.

B Belief Quality Metrics

We define six metrics to assess the fidelity of predicted belief states on AgentClinic episodes. All metrics are computed per-turn and averaged across episodes.

Evidence F1. Micro-F1 between the model's predicted evidence set $\hat{E}_t = \texttt{confirmed}_t \cup \texttt{denied}_t$ and the environment's ground-truth disclosed findings $\mathcal{R}_{\leq t}$ (cumulative up to turn t). Matching uses UMLS concept IDs (CUI) with exact-match fallback on normalized surface forms.

Assessment Accuracy. The fraction of turns where `assessment.top` matches the scenario's ground-truth diagnosis. Matching uses a curated synonym table (e.g., "MI" = "Myocardial Infarction") plus LLM-based equivalence check for ambiguous cases.

Next-step Hit. For each turn t, we check whether at least one question in `next_step.critical` is semantically matched (BERTScore ≥ 0.85) to the actual doctor utterance in turns $t{+}1$ or $t{+}2$. The metric is the fraction of turns where a hit occurs.

Non-Contradiction. Let C_t = `confirmed` at turn t and D_t = `denied` at turn t. A contradiction occurs when a CUI appears in C_{t-1} and D_t without the patient having explicitly negated it in turn t's response. Non-Contradiction $= 1 -$ (contradictions$/|C_{t-1}|$), averaged across turns and episodes.

Question Efficiency. The fraction of doctor question turns that produce at least one new verified evidence item in the subsequent patient response: $|\mathcal{E}^{\text{new}}_{t+1} \cap \mathcal{R}_{t+1}| \geq 1$.

Confidence Calibration. We partition all turns by their `confidence` level and compute diagnosis accuracy within each group. Calibration is defined as $\text{Acc}(\textsc{High}) - \text{Acc}(\textsc{Low})$, normalized to $[0, 1]$. A higher value indicates the model's confidence is discriminative of actual diagnostic correctness.

C RL Training Details

C.1 Patient Simulator

Stage 2 uses a frozen Qwen2.5-72B-Instruct as the patient agent. The scenario pool consists of 2,400 unique clinical cases constructed from MedDialog-EN.

Scenario construction. Each MedDialog-EN entry provides a patient description (chief complaint, brief history) and a doctor response (which often suggests differential diagnoses and recommended tests). We construct full clinical profiles in two steps: (1) we extract all clinical facts explicitly mentioned in both the patient message and doctor response using MedSpaCy; (2) we prompt GPT-4o to supplement additional findings that are medically consistent with the identified diagnosis, drawing on standard clinical presentations for that condition. Each final profile contains a ground-truth diagnosis, positive findings, pertinent negatives that dis-

tinguish from plausible differentials, and relevant history.

Scenario diversity. Cases span 18 medical specialties (internal medicine, neurology, cardiology, etc.) with no overlap with evaluation benchmarks (HealthBench, AgentClinic, MedQA, PubMedQA, MedMCQA, MMLU). Overlap is verified by matching ground-truth diagnoses and chief complaints against benchmark case lists.

C.2 Profile-Grounded Fact Verification ($\mathcal{R}_t$)

The belief-progress reward requires determining which clinical facts the patient disclosed at each turn. Since each simulated patient is instantiated from a structured clinical profile, the environment verifies disclosures directly against this profile.

Profile fact structure. Each RL training profile enumerates its clinical facts at the individual-finding granularity. Each fact $f \in \mathcal{F}$ has a canonical name (e.g., "constant vertigo") and an assertion (`confirmed` or `denied`). Facts correspond to independently verifiable clinical observations: "substernal chest pain" and "radiation to left arm" are separate facts.

NLI alignment model. We use `PubMedBERT-MNLI-MedNLI`, a PubMedBERT model fine-tuned on MultiNLI then MedNLI (Romanov and Shivade, 2018). Given a profile fact as premise and a text segment as hypothesis, the model outputs ENTAILMENT (fact disclosed), CONTRADICTION (fact denied), or NEUTRAL (not addressed). The model is frozen throughout RL training and does not participate in policy optimization.

$\mathcal{R}_t$ construction. After each patient response p_t, the environment applies the NLI model to every (sentence, profile fact) pair. A fact $f \in \mathcal{F}$ is added to $\mathcal{R}_t^+$ if some sentence s in p_t entails f, and to $\mathcal{R}_t^-$ if some sentence contradicts f. The reward uses $\mathcal{R}_t = \mathcal{R}_t^+$ for the positive evidence term. Denied facts in $\mathcal{R}_t^-$ contribute to the contradiction penalty when the agent incorrectly adds them to `confirmed`. The environment maintains a cumulative set $\mathcal{D}_{\leq t} = \bigcup_{i=1}^{t} \mathcal{R}_i^+$ for redundancy detection.

Belief-to-fact alignment. The same NLI model aligns agent belief findings to profile facts. For each new finding $e \in \mathcal{E}_t^{\text{new}}$, the model evaluates $\text{NLI}(f, e)$ for all $f \in \mathcal{F}$. Finding e receives positive reward only if it maps (via entailment) to some fact in $\mathcal{R}_t$.

C.3 Reward Verification Audit

The belief-progress reward relies on automatic profile-grounded verification rather than an LLM judge. To validate this pipeline, we audit whether its decisions agree with human judgments.

Sampling. We randomly sample 300 turns from held-out RL rollouts, stratified by reward sign (positive / zero / negative, 100 each) and dialogue stage (early / middle / late). This yields 2,847 response–fact pairs for disclosure detection and 412 belief–fact pairs for alignment evaluation.

Annotation. Two clinically-trained annotators independently label each item. For disclosure detection, annotators judge whether each profile fact is entailed, contradicted, or not mentioned in the patient response. For belief alignment, they judge whether the agent's newly confirmed finding correctly maps to the matched profile fact. Inter-annotator agreement is $\kappa = 0.83$; disagreements are adjudicated by a third reviewer.

Results. Table 7 reports pipeline–human agreement. Positive fact disclosure achieves F1 = 92.3%, confirming that the primary reward signal is reliable. Denied-fact detection is lower (F1 = 87.8%); however, denied facts only trigger the contradiction penalty under strict conditions, limiting the impact on training. Overall reward-sign agreement reaches $\kappa = 0.82$, indicating the fixed pipeline provides a sufficiently reliable training signal for online RL.

Table 7: Human audit of the reward verification pipeline on 300 stratified RL turns. Two clinically-trained annotators ($\kappa_{\text{IAA}} = 0.83$).

Component	Metric	Result
Positive fact disclosure ($\mathcal{R}_t^+$)	F1	92.3
Denied fact disclosure ($\mathcal{R}_t^-$)	F1	87.8
Belief-to-fact alignment	Accuracy	91.5
Redundancy detection	F1	94.9
Contradiction detection	F1	85.9
Overall reward-sign agreement	κ	0.82

C.4 PPO Training Details

We use PPO with Stage 1 data replay ($\sim$15% of each batch) to prevent format degradation, and adaptive KL penalty (β_{KL}, target=0.02) to constrain policy drift. The per-turn reward is $r_t =$

r_t^{prog}; the terminal reward r_T is added at the final turn.

Table 8: PPO hyperparameters for Stage 2 training.

Parameter	Value
Learning rate	1×10^{-5} (cosine decay)
Batch size	128 episodes
Mini-batch size	16
PPO clip ratio	0.2
GAE λ	0.95
Discount γ_{RL}	0.99
KL target	0.02
KL penalty coefficient	adaptive
Stage 1 replay ratio	15%
Max turns per episode ($T_{\max}$)	20
Total training episodes	50k
Value function coefficient	0.5
Entropy bonus	0.01
LoRA rank (r)	64

D Evaluation Protocol

We evaluate all models under the official protocols of each benchmark to ensure reproducibility. Below we list system prompts, generation parameters, and scoring mechanisms.

D.1 HealthBench

All models receive the default system prompt from the OpenAI `simple-evals` framework:

```
You are a helpful assistant.
```

Each model generates a single completion for the full multi-turn conversation history using the official default parameters (temperature=0.5, max tokens=2048, temperature sampling).

Grading. Each response is scored against physician-authored rubric items by an LLM judge (GPT-4.1). For each rubric item the judge returns a binary `criteria_met` label; the overall score is $\sum$ achieved points / $\sum$ positive points, clipped to $[0, 1]$. Rubric content and scoring logic follow the official HealthBench release (Arora et al., 2025).

D.2 AgentClinic

AgentClinic simulates an OSCE-style multi-turn consultation with four agents. All models share identical prompt templates.

Doctor Agent Prompt

You are a doctor named Dr. Agent who only responds in the form of dialogue. You are inspecting a patient who you will ask questions in order to understand their disease. You are only allowed to ask `{MAX_INFS}` questions total before you must make a decision. You have asked `{infs}` questions so far. You can request test results using the format "`REQUEST TEST: [test]`". Your dialogue will only be 1–3 sentences in length. Once you have decided to make a diagnosis please type "`DIAGNOSIS READY: [diagnosis here]`".

Patient Agent Prompt

You are a patient in a clinic who only responds in the form of dialogue. You are being inspected by a doctor who will ask you questions and will perform exams on you in order to understand your disease. Your answer will only be 1–3 sentences in length.

The patient agent is parameterized with the full scenario profile (demographics, symptoms, history) and instructed not to reveal the diagnosis explicitly.

Moderator (Diagnosis Judge)

You are responsible for determining if the correct diagnosis and the doctor diagnosis are the same disease. Please respond only with Yes or No.

Interaction parameters. All parameters follow the official AgentClinic release (Schmidgall et al., 2026): max turns=20, temperature=0.05, max tokens per response=200. The patient agent, measurement agent, and moderator are fixed across all evaluated doctor models.

E Case Study: Comparative Dialogue Analysis

Below we show an AgentClinic case (scenario 71) where the same patient interacts with three different doctor agents. Ground-truth diagnosis: **Labyrinthitis** (post-viral inner ear inflammation). Doctor-R1 diagnoses incorrectly (✗); DoctorAgent-RL diagnoses incorrectly (✗); BridgeGap diagnoses correctly (✓).

Presenting Complaint (shared by all agents)

"I've been feeling like the room is spinning since this morning, and it's been constant. I also feel a bit nauseous at times."

Doctor-R1 — Final diagnosis: BPPV ✗

Turn 1 Doctor: Is the dizziness constant or does it come and go?
Patient: It's constant, like the room is spinning. Started this morning and hasn't let up.

Turn 2 Doctor: Have you experienced any hearing loss, tinnitus, or ringing in the ears?
Patient: No hearing loss or ringing. Just the spinning and occasional nausea.

Turn 4 Doctor: Does the dizziness worsen with certain head movements?
Patient: A bit worse when I turn my head quickly, but it's mostly just constant. [constant ≠ BPPV]

Turn 5 Doctor: REQUEST TEST: Vestibular_Function_Test
Result: NORMAL READINGS

Turn 6 Doctor: Have you had any recent viral infections, such as a cold?
Patient: Yes, I had a cold a few days ago. [key clue—not synthesized]

Turn 8 Doctor: Any neck pain or cervical spine issues?
Patient: No. [low-yield]

Turn 9 Doctor: Any changes in diet, fluid intake, or electrolyte balance?
Patient: I've been drinking plenty of fluids. Only change is extra garlic. [low-yield]

Turn 17 Doctor: DIAGNOSIS: Benign paroxysmal positional vertigo (BPPV) with possible viral etiology. ✗
▷ Contradictory: BPPV is episodic/positional, but patient's vertigo is constant. Viral precursor supports labyrinthitis, not BPPV.

Doctor-R1 collects all discriminative findings—constant vertigo (Turn 1), recent URI (Turn 6)—but fails to synthesize them. It diagnoses BPPV despite the patient describing *constant* vertigo, and acknowledges a viral precursor that contradicts BPPV.

DoctorAgent-RL — Final diagnosis: "Vertigo" ✗

Turn 1 Doctor: How can I assist you today?
Patient: I've been feeling like the room is spinning since this morning. Also a bit nauseous.

Turn 2 Doctor: Do you have any history of dizziness or vertigo?
Patient: No, I've never had anything like this before. Constant, low-grade spinning all day.

Turn 3 Doctor: Have you noticed any changes in vision or hearing?
Patient: No, vision and hearing seem normal. The spinning is constant. [no follow-up on triggers]

Turn 4 Doctor: Based on your symptoms, you may be suffering from vertigo, possibly related to a viral infection...
Patient: I'll follow your advice.

Turn 5 Doctor: DIAGNOSIS: Vertigo, possibly viral. ✗ [correct direction, but non-specific]

DoctorAgent-RL intuits the viral link but stops at a vague "vertigo" label without specifying the etiology, performing no differential reasoning between BPPV, labyrinthitis, and vestibular neuritis.

BridgeGap — Final diagnosis: Labyrinthitis ✓

Turn 1 Doctor: Can you describe the dizziness—is it a constant spinning, or brief episodes triggered by head movement?
Patient: It's constant, room is spinning since this morning. Hasn't stopped.
▷ Belief: top = Peripheral vertigo; conf = Low; confirmed: acute_onset, constant_vertigo

Turn 2 Doctor: Have you had a recent cold, flu, or ear infection in the past week?
Patient: Yes, I had a cold a few days ago.
▷ Belief: top = **Labyrinthitis**; conf = Med; +1 recent_URI; alts: Vestib. neuritis, BPPV

Turn 3 Doctor: Have you noticed any hearing loss or ringing in your ears? Any ear fullness or pressure?
Patient: No hearing loss or ringing. Just spinning and nausea.
▷ Belief: conf = Med; denied: hearing_loss; BPPV dropped (vertigo is constant, not positional)

Turn 4 Doctor: Any numbness, weakness, slurred speech, or severe headache? Does the vertigo change with position?
Patient: No numbness or weakness. Slightly worse with quick head turns but mostly constant.
▷ Belief: conf = **High**; denied: neuro_sx; central causes excluded; no critical questions remain

Turn 5 Doctor: DIAGNOSIS READY: Labyrinthitis—acute vestibular inflammation following a recent upper respiratory infection. ✓
▷ Belief: action = DIAGNOSE

Summary. This case illustrates the mechanism behind the aggregate results in §4. Doctor-R1 collects all discriminative findings—constant vertigo, recent URI, absence of focal neurological signs—but lacks a structured representation to synthesize them, yielding a contradictory BPPV diagnosis. DoctorAgent-RL intuits the viral link but terminates with a non-specific "vertigo" label without differential reasoning. BridgeGap's belief state explicitly tracks these three findings, uses `next_step.critical` to prioritize the most discriminative question at each turn, and reaches the correct, specific diagnosis (labyrinthitis) in 5 turns—explaining its higher diagnostic accuracy and earlier stopping behavior in the quantitative evaluation.