

Ask Better Questions: Learning Efficient Information-Seeking Policy for Clinical Cold-Start Dialogue

Li Wenbo^{1,*} Le Cai^{1,*} Jiawei Liu¹ Jinhong Yu²
Zhuo Chen¹ Zhuoran Luo¹ Yang Qingqing³ Gao Cheng³
Qikai Cheng¹ Wei Lu¹ Xiaozhong Liu^{2,†}

¹Wuhan University ²Worcester Polytechnic Institute

³The Second Affiliated Hospital of Wannan Medical College

*Equal contribution.

†Corresponding author.

Abstract

Despite strong performance in medical question answering, large language models remain limited in clinical cold-start settings, where little patient information is initially available and key evidence must be acquired through dialogue. This requires models to decide what to ask under a limited turn budget, balancing broad history taking with targeted clarification of high-risk cues. To address the aforementioned challenges, we propose MedCIA, a two-stage framework for improving clinical information acquisition. First, it reconstructs patient records into structured inquiry trajectories, enabling supervised learning of expert-guided information-seeking intents and the breadth-depth trade-off in clinical questioning. Second, it refines the policy with Adaptive Tree Policy Optimization (ATPO), an uncertainty-aware reinforcement learning procedure that allocates rollout computation to ambiguous dialogue states in a profile-anchored patient simulator. Furthermore, we curate a multi-turn dataset (MedIntake) for active clinical intake with progressive patient information disclosure and diverse inquiry strategies. Experiments on an EHR-derived benchmark show that MEDCIA improves importance-weighted coverage, early acquisition, and cost control under controlled simulation.

1 Introduction

Large language models (LLMs) are moving from static medical question answering to interactive clinical assistance. Prior work reports expert-level or near-expert-level performance on medical QA benchmarks (Singhal et al., 2025), and recent randomized trials show that patient-facing LLM chatbots can support upstream workflows by collecting pre-consultation information (Tao et al., 2026).

In medical dialogue, clinical cold start refers to settings where little prior patient information is available and critical evidence must be acquired

through limited interaction. Existing multi-turn medical agents formulate consultation as reinforcement learning over simulated patients (Feng et al., 2025), convert medical notes into history-taking dialogues (Zhou et al., 2026), or optimize conversational quality and medical decision-making (Lai et al., 2025). However, they do not directly target patient-background recovery under strict turn budgets. This setting is especially relevant to older-adult and chronic-care intake. Older patients often present with dense longitudinal backgrounds, such as multimorbidity, polypharmacy, prior procedures, allergies, and social-care constraints, yet an LLM-facing intake system may still start with little accessible prior information (Watson et al., 2025; National Institute for Health and Care Excellence, 2016; American Geriatrics Society Beers Criteria® Update Expert Panel, 2023). Moreover, long model-led interactions may be impractical for some older adults due to digital literacy, cognitive or physical burden, usability barriers, and caregiver-mediated communication (Hepburn et al., 2025). This motivates recovering high-value patient-reportable evidence within a few turns.

Many diagnosis-oriented multi-turn medical agents are evaluated using outcome-oriented metrics such as diagnostic accuracy, exact-match diagnosis, or top- k diagnostic recall. Yet information acquisition and diagnostic decision-making are clinically distinct: final diagnosis and treatment should remain under qualified clinician oversight, while patient-facing LLMs may be better suited to collecting and surfacing clinically relevant evidence. We therefore study whether an LLM can efficiently uncover key patient-reportable evidence, including symptoms, chronic conditions, medications, allergies, prior procedures, family history, and social history, under sparse information and tight turn budgets.

We introduce clinical cold-start information acquisition: given minimal initial patient information,

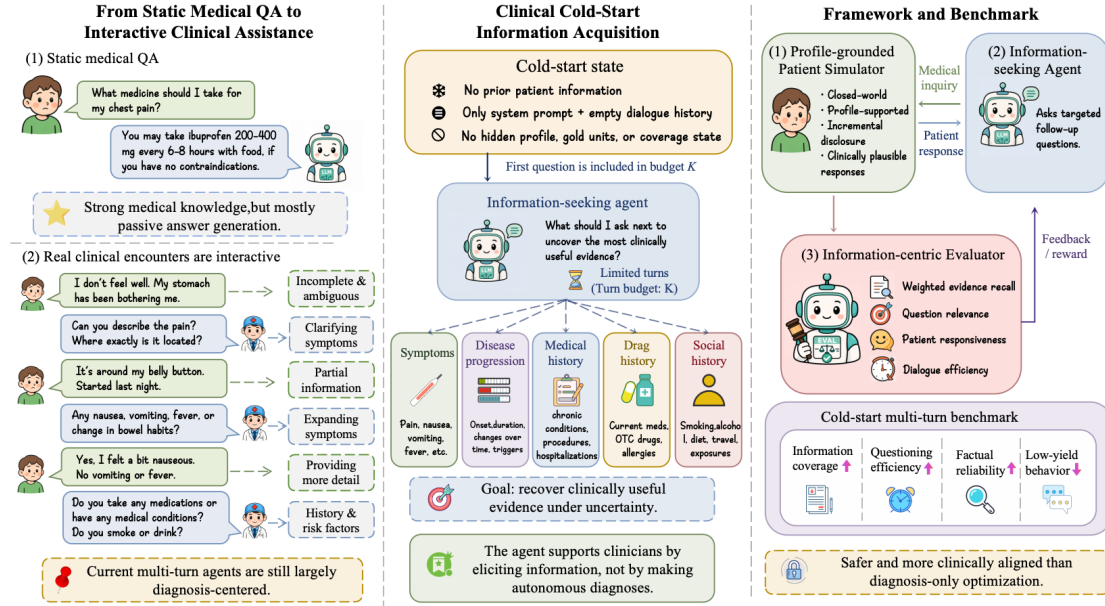


Figure 1: From static medical QA to clinical cold-start information acquisition. The figure contrasts passive answer generation and diagnosis-centered interaction with a limited-turn intake setting, where an information-seeking agent starts without prior patient information and is evaluated through a profile-grounded simulator, information-centric evaluator, and benchmark metrics for evidence coverage, efficiency, reliability, and low-yield behavior.

an agent chooses questions within a constrained dialogue budget to maximize clinically useful evidence while leaving final diagnostic judgment to clinicians. This formulation requires more than fluent medical conversation. A strong intake policy must recover distributed background, identify when risk-relevant cues deserve targeted follow-up, and avoid wasting limited turns on redundant or low-yield questions.

To address this problem, we propose MEDCIA, a two-stage framework that reconstructs patient records into structured inquiry trajectories for supervised fine-tuning and adapts ATPO for utility-aware policy refinement. We further introduce MEDINTAKE, a profile-grounded benchmark for limited-turn active information acquisition. To bridge the gap between simulation and clinical intuition, we curate a clinician-adjudicated reference set with blind consultations and post-hoc physician revisions. Although resource-intensive, these expert annotations calibrate breadth–depth questioning behavior. Experiments show that MEDCIA improves importance-weighted coverage, early acquisition, factual reliability, and low-yield behavior under controlled simulation.

Our contributions are summarized as follows:

- We formulate clinical cold-start information acquisition as a cost-sensitive conversational evidence elicitation problem, shifting

multi-turn medical dialogue from diagnosis-centered prediction to clinician-assistive evidence elicitation.

- We propose **MedCIA**, a two-stage framework that combines profile-grounded trajectory reconstruction with utility-aware policy refinement to elicit high-value patient-reportable evidence under limited turns.
- We introduce **MEDINTAKE**, a profile-grounded multi-turn clinical intake benchmark derived from EHR records, with a unified evaluation protocol covering weighted atomic coverage, acquisition efficiency, factual reliability, and low-yield behavior.¹

2 Related Work

Clinical information acquisition and diagnostic dialogue. Prior work often formulates medical history taking as structured completion or prediction, decomposing dialogue into NLU, dialogue state

¹Due to MIMIC data-use restrictions, patient-specific profiles, dialogue trajectories, gold units, or profile-level splits will not be released through unrestricted platforms. We will release non-patient-specific artifacts, including schemas, prompts, evaluation code, metric implementations, synthetic examples, documentation, and reconstruction scripts for credentialed MIMIC users, and will explore a PhysioNet credentialed-access release for the processed benchmark.

tracking, and slot filling, and evaluating field coverage or symptom classification accuracy, as in MediTOD (Saley et al., 2024). Diagnosis-oriented systems further model inquiry as sequential symptom selection optimized for diagnostic outcomes, using structured disease–symptom relations, discrete action spaces, or hierarchical policies (Wei et al., 2018; Xu et al., 2019; Chen et al., 2022; Zhong et al., 2022). These formulations work well in closed symptom spaces, but often center on exhaustive completion or diagnostic prediction. Related work in conversational recommendation has similarly explored interactive preference clarification, including maximum-entropy-based strategies for eliciting informative user feedback (Zhu et al., 2025). When diagnosis remains under clinician oversight, the key challenge for patient-facing LLMs is prioritizing high-yield patient-reportable evidence within limited turns.

Clinical-note-to-dialogue construction. Because real patient–clinician dialogues are costly and privacy-sensitive, prior work constructs dialogue resources from notes, case records, and simulated consultations. PriMock57 (Papadopoulos Korfatis et al., 2022), ACI-BENCH (Yim et al., 2023), and MTS-Dialog (Ben Abacha et al., 2023) mainly support documentation generation, dialogue summarization, or ASR. NoteChat (Wang et al., 2024), MedSynth (Mianroodi et al., 2025), and Note2Chat (Zhou et al., 2026) generate patient–clinician dialogues from notes or cases for history taking and diagnostic reasoning. However, these note-centered frameworks are typically organized around chief complaints, HPI reconstruction, documentation, or diagnostic reasoning, and lack explicit ground-truth profiles with structured evidence units for fine-grained recovery evaluation.

LLM-based clinical agents and policy learning. Recent LLM-based clinical dialogue systems move beyond static QA toward interactive consultation. AMIE (Tu et al., 2025) improves conversational diagnostic capability through self-play and clinician-guided evaluation, while DoctorAgent-RL (Feng et al., 2025) formulates medical consultation as dynamic decision-making under uncertainty. Other work aligns diagnostic reasoning with clinician preferences (Xu et al., 2024), incorporates EHR-guided policy shaping (Fang et al., 2026), explores ensemble reasoning over structured medical knowledge for clinical decision support (Rashid and McGuinness, 2025), or develops virtual clinical

environments for policy refinement (Qiu et al., 2025; Reis et al., 2025). These systems improve interaction quality, reasoning, or diagnosis-oriented decision-making, but do not directly optimize structured patient-profile evidence recovery when conversational turns are costly and strictly constrained.

Evaluation of medical dialogue. Existing dialogue datasets and benchmarks, including MedDialog (Zeng et al., 2020), MedDG (Liu et al., 2022), MediTOD (Saley et al., 2024), simulated interview data (Fareez et al., 2022), and MedSynth (Mianroodi et al., 2025), primarily emphasize response quality, symptom-level coverage, documentation quality, or final diagnostic accuracy; complementary work has also developed user-centered evaluation frameworks for healthcare conversations (Abbasian et al., 2024). While these objectives are valuable for general consultation and clinical dialogue evaluation, they are not designed specifically to measure strategic, turn-efficient recovery of longitudinal patient evidence. To bridge this gap, we introduce MEDINTAKE, a benchmark that formalizes cold-start clinical intake as cost-sensitive information acquisition. Under a unified, profile-grounded protocol, MEDINTAKE evaluates conversational agents across four dimensions: importance-weighted coverage, acquisition efficiency, factual reliability, and low-yield behavior.

3 Methodology

We propose **MEDCIA**, a cost-sensitive framework for clinical cold-start information acquisition. Figure 2 summarizes its four components: patient profile construction, profile-anchored dialogue corpus construction, expert-guided SFT trajectories, and utility-aware policy refinement.

3.1 Patient Profile and Dialogue Corpus Construction

We construct structured patient profiles from processed MIMIC-IV and MIMIC-IV-Note records (Johnson et al., 2023b,a), using discharge summaries as the primary source and structured encounter metadata as complementary evidence. Multiple encounters for the same patient are chronologically aggregated to preserve longitudinal history, including chronic conditions, long-term medications, allergies, procedures or hospitalizations, family history, social history, and current presentation cues. LLM-assisted post-processing normalizes raw notes into patient-reportable

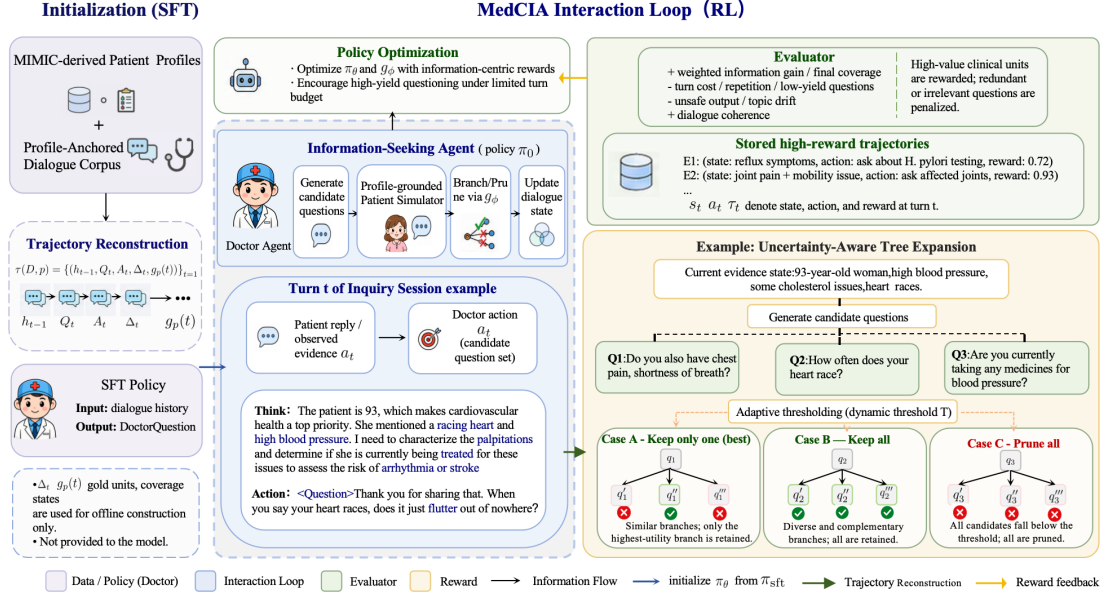


Figure 2: Overview of the MedCIA framework for clinical cold-start information acquisition.

fields while filtering post-hoc clinician-only facts. Temporal aggregation, leakage control, and release constraints are detailed in Appendix A.1.

From these profiles, we construct a profile-anchored dialogue corpus. A doctor agent asks one focused question per turn, and a closed-world patient simulator incrementally reveals only profile-supported facts. For scalable supervision, a trajectory generator scores candidate questions by expected weighted coverage gain, contextual relevance, timing appropriateness, and redundancy control; this oracle-assisted signal is used only for offline data construction and never at validation or test time. We also construct 80 clinician-revised reference trajectories from profile-blind consultations followed by post-hoc physician revision. These trajectories support quality control, calibration, and qualitative analysis of breadth–depth questioning. Additional details, statistics, and examples are provided in Appendix A.2 and Appendix A.1.

3.2 Trajectory Reconstruction and Supervised Fine-Tuning

Raw transcripts contain question–answer sequences but not the clinical evidence acquired at each turn. Direct imitation may therefore overfit surface wording or turn order rather than high-yield information-seeking structure. We reconstruct each dialogue into a structured trajectory that links every turn to the newly elicited clinical information.

Atomic units and weighted coverage. For each

patient case p , we decompose the hidden profile into a set of atomic medical information units $\mathcal{U}_p = \{u_1, \dots, u_{N_p}\}$, which serve as the acquisition targets of the intake dialogue. Each unit $u \in \mathcal{U}_p$ is assigned a deterministic evidence-based importance weight

$$w_p(u) = \Gamma(\phi_p(u); \mathcal{E}) \quad (1)$$

where $\phi_p(u)$ denotes the normalized attributes of the unit, including its clinical category, risk attribute, and evidence-table match. $\mathcal{E}$ is a frozen weighting specification shared across all evaluated models. For unmatched units, we assign a documented category-level prior and record this fallback through `weight_source_ref`.

Given the dialogue history h_t after turn t , an offline fixed evaluator extracts candidate elicited units $\hat{\mathcal{U}}_t$ from the dialogue and matches them against the gold units in $\mathcal{U}_p$. This produces the covered unit set $\mathcal{C}_t \subseteq \mathcal{U}_p$ and the newly covered set $\Delta_t = \mathcal{C}_t \setminus \mathcal{C}_{t-1}$, with $\mathcal{C}_0 = \emptyset$.

To express evidence acquisition as a unified set utility, we define

$$F_p(\mathcal{C}) = \frac{1}{Z_p} \sum_{u \in \mathcal{C}} w_p(u) \quad Z_p = \sum_{u \in \mathcal{U}_p} w_p(u) \quad (2)$$

where Z_p denotes the total importance mass of case p . The importance-weighted coverage and the incremental weighted gain are then given by

$$\begin{aligned} \text{wCov}_p(t) &= F_p(\mathcal{C}_t) \\ g_p(t) &= F_p(\mathcal{C}_t) - F_p(\mathcal{C}_{t-1}) \end{aligned} \quad (3)$$

Weighting sources, matching rules, and a clinician face-validity check of atomic-unit definitions and importance bins are provided in Appendix C.1 and Appendix D.2.

Trajectory reconstruction and SFT objective. Given a dialogue $D = \{(Q_t, A_t)\}_{t=1}^T$ and its hidden profile p , we reconstruct the dialogue into a structured information-acquisition trajectory:

$$\tau(D, p) = \{(h_{t-1}, Q_t, A_t, \Delta_t, g_p(t))\}_{t=1}^T \quad (4)$$

Here, h_{t-1} is the pre-turn dialogue history, Δ_t the newly elicited atomic units, and $g_p(t)$ the incremental weighted gain. These annotations make the dialogue’s information-seeking structure explicit for supervision analysis.

During supervised fine-tuning, the model observes only the preceding dialogue history and generates the next focused doctor question:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(D,p) \sim \mathcal{D}_{\text{SFT}}} \left[\sum_{t=1}^T \log \pi_{\theta}(Q_t \mid h_{t-1}) \right] \quad (5)$$

Here, $\pi_{\theta}(Q_t \mid h_{t-1})$ denotes the sequence likelihood of the target doctor question. The target output is restricted to a single `<DoctorQuestion>` field. At inference time, the model receives only the system prompt and dialogue history; hidden profiles, gold units, coverage states, gap lists, evaluator signals, and offline annotations are not provided. Trajectory construction and token-level likelihood details appear in Appendix B.1.

3.3 Reinforcement Learning for Cost-Sensitive Information Acquisition

Interactive acquisition environment. SFT provides a stable initialization but does not directly optimize long-horizon acquisition under a strict question budget. We formulate cold-start intake as a partially observable, cost-sensitive sequential decision problem: the hidden profile defines recoverable patient-reportable evidence, while the doctor policy observes only dialogue history. The environment contains hidden profiles $\mathcal{P}$, histories $\mathcal{H}$, doctor-question actions $\mathcal{A}$, an information-centric reward $\mathcal{R}$, and maximum rollout length T_{max} . At each turn, the policy asks one focused question, receives a profile-grounded response, and is scored by the resulting acquisition state.

Uncertainty-aware tree policy refinement. We adapt ATPO (Cao et al., 2026) to profile-grounded

clinical intake. At dialogue state x_k , the doctor policy samples N candidate macro-actions $\{y_k^i\}_{i=1}^N$, each a complete structured doctor question. Each candidate is passed to the profile-grounded patient simulator, which generates a simulated response and successor state, enabling comparison across alternative inquiry paths.

Following ATPO, expansion is guided by a one-step value estimate and an uncertainty score:

$$\begin{aligned} \hat{Q}_{\psi}(x_k, y_k^i) &= r_k^i + \gamma V_{\psi}(x_{k+1}^i), \\ U(x_k) &= \alpha_U U_1(x_k) + (1 - \alpha_U) U_2(x_k) \end{aligned} \quad (6)$$

Here, U_1 captures critic-based value uncertainty and U_2 candidate-value variance. Higher uncertainty allocates more rollouts to broader question-sequence exploration.

Information-centric reward. The reward targets evidence acquisition rather than diagnostic correctness. At the turn level, we use

$$\begin{aligned} r_k &= \boldsymbol{\lambda}^{\top} \mathbf{s}_k, \\ \mathbf{s}_k &= [\text{Fmt}_k, \text{Ver}_k, \text{Eff}_k]^{\top} \end{aligned} \quad (7)$$

where Fmt_k rewards the required `DoctorQuestion` format, Ver_k checks the patient-reportable boundary, and Eff_k measures question effectiveness.

At the branch level, the terminal reward optimizes recovered importance-weighted evidence:

$$\begin{aligned} \text{WRecall}_p(D_{1:T}) &= F_p(\mathcal{C}_T) \\ R_T &= \beta_g \text{WRecall}_p(D_{1:T}) \end{aligned} \quad (8)$$

Here, $\text{WRecall}_p(D_{1:T})$ is the weighted recall of recovered gold units, and β_g controls terminal reward strength.

After expansion, root-to-leaf branches are converted into training trajectories through value traceback. The policy is updated with a clipped token-level objective, and the critic predicts traceback values. Scorer coefficients, traceback equations, invalid-output handling, inactive reward terms, and hyperparameters appear in Appendix B.3 and Appendix B.2.

4 Model Training and Evaluation

This section reports the training and evaluation setup, main results, and ablations. MEDCIA uses trajectory-reconstruction SFT followed by utility-aware policy refinement with adapted ATPO; objectives, reward implementation, and hyperparameters are provided in Appendix B.1, Appendix B.3, and Appendix B.2.

4.1 MEDINTAKE: Construction and Split

We construct MEDINTAKE from structured profiles derived from credentialed-access MIMIC clinical resources (Johnson et al., 2023a,b). The corpus contains 2,400 patient profiles and enables evaluation beyond single-encounter symptom elicitation through longitudinal background fields such as chronic conditions, medications, procedures, allergies, family history, and social history.

We use profile-level 2,000/200/200 train/dev/test splits with no profile overlap between training and evaluation. Beyond large-scale simulated trajectories for scalable turn-level supervision, MEDINTAKE includes 80 clinician-revised reference trajectories paired with profile-blind consultations for quality control and trajectory calibration. Detailed corpus statistics, validity controls, and training configuration are provided in Appendix A, Appendix A.1, and Appendix B.2.

The 200 test cases contain 5,183 gold atomic units, with an average of 25.9 units per case. Medical history accounts for 37.4% of the units, medications for 25.5%, demographics for 11.5%, allergies for 9.0%, family history for 7.1%, and social history for 4.4%, with the remaining categories accounting for 5.1%. A full audit shows that 96.5% of the test-set gold units belong to categories typically reportable by patients or caregivers, and excluding the remaining 3.5% does not change the model ranking. Each training profile retains between 1 and 128 simulated trajectories after tree expansion and filtering. For quality control, two annotators evaluated simulated dialogues from 100 cases: the patient simulator achieved mean scores of 4.73/5 for naturalness and 4.95/5 for profile faithfulness, while 97.3% of responses contained no fabricated or contradictory facts. Only 6.8% of traceable information was disclosed in the first turn, with a mean first-appearance turn of 6.97.

4.2 Training Procedure

We train MEDCIA in two stages with Qwen3-8B (Yang et al., 2025) as the backbone: trajectory-reconstruction SFT followed by closed-loop policy refinement in a profile-grounded simulator.

Trajectory-reconstruction supervised fine-tuning. Each SFT instance is derived from a reconstructed acquisition trajectory, but the model-visible input contains only the system prompt and dialogue history. The target is a single `<DoctorQuestion>`, encouraging one focused question per turn rather

than reasoning traces or multi-question outputs. Hidden profiles, gold units, coverage states, scores, and trajectory annotations are used only offline.

Utility-aware policy refinement. Starting from the SFT checkpoint, candidate doctor questions are expanded in an uncertainty-aware dialogue tree. The patient simulator follows closed-world, incremental-disclosure constraints, and each branch is scored by format compliance, boundary verification, question effectiveness, and terminal importance-weighted recall. Backward traceback computes target values; the policy is then updated with a clipped token-level objective, and the critic predicts traceback values. Details are provided in Appendix B.3 and Appendix B.2.

4.3 Evaluation Protocol and Metrics

All models are evaluated in the same cold-start environment. For each case, the model interacts with a profile-anchored simulated patient for at most $K = 15$ doctor turns, with one focused question per turn; the fixed cold-start opening question counts toward this budget. The model receives only dialogue history and patient responses, without hidden profiles, gold units, weights, coverage states, gap lists, or evaluator outputs. A fixed offline evaluator extracts elicited units, matches them to hidden gold units, computes metrics, and applies the same early-stopping rule across systems.

We report final weighted coverage (wCov@15), early weighted coverage (wEarly@5), acquisition speed (wAUC@15), unweighted breadth (Cov@15), factual reliability, and low-yield behavior. Full formulas, denominator definitions, carry-forward alignment, supplementary metrics, and normalization rules are provided in Appendix D.1 and Appendix D.2.

Evaluator, simulator, and prompt sensitivity. The main results use a fixed Qwen3.6-Flash evaluator and a profile-grounded patient simulator, both deployed locally. Appendix D.2 reports detailed evaluator- and simulator-sensitivity results. Re-evaluating the same model outputs and patient responses with locally deployed Qwen-, DeepSeek-, and Llama-family evaluators yields MEDCIA-GPT-5.4 wCov@15 differences of +10.4, +10.1, and +7.5 points, respectively. We also evaluate 13 GPT-5.4 prompt variants on the same 10-case subset; the strongest GPT-5.4 configuration reaches 60.62 wCov@15, compared with 69.53 for MEDCIA on the same cases. We additionally vary the

Table 1: Main results on cold-start clinical information acquisition. All models are evaluated on the 200-case test split under a maximum dialogue budget of 15 doctor questions, with evaluator-driven early stopping after turn 12. Percentage-valued metrics are reported on a 0–100 scale. $\uparrow$ indicates higher is better, while $\downarrow$ indicates lower is better.

Model	wCov@15 $\uparrow$	wAUC@15 $\uparrow$	wEarly@5 $\uparrow$	Cov@15 $\uparrow$	Prec. $\uparrow$	Halluc. Rate $\downarrow$	Contrad. Rate $\downarrow$	Low-Yield Rate $\downarrow$
<i>Closed-source model</i>								
GPT-5.4	59.3	41.8	33.6	56.5	97.5	2.5	0.3	45.4
Qwen3.6-Max-Preview	50.7	38.9	34.0	45.3	73.4	26.6	0.8	65.1
DeepSeek-V4-Flash	58.3	44.7	39.9	52.5	81.9	18.1	0.7	58.2
Baichuan-M3-Plus	38.1	32.7	31.3	33.8	86.3	13.7	0.4	83.6
<i>Open-source models and task-related adapted baselines</i>								
MedGemma-4B-it	35.7	25.8	22.0	33.8	90.8	9.2	0.1	76.4
MedGemma-27B-text-it	41.2	32.9	30.2	38.3	75.5	24.5	0.4	72.3
DoctorAgent-RL	42.5	31.6	27.0	38.7	81.0	19.0	0.3	69.6
Note2Chat-MT	43.7	31.5	27.3	40.1	82.2	20.3	0.9	63.2
<i>Profile-grounded information acquisition model</i>								
MEDCIA	69.7	47.1	40.3	64.3	89.8	10.2	0.2	45.1

Note. wCov@15 denotes final importance-weighted atomic coverage; wAUC@15 is the area under the weighted coverage curve; wEarly@5 measures early weighted coverage after the first five doctor questions; Cov@15 denotes unweighted atomic coverage. Prec. and Halluc. Rate measure factual precision and unsupported confirmed positive assertions under the same positive-fact denominator, while Contrad. Rate is computed separately over gold atomic units. Therefore, these columns are fact-level reliability rates with different denominators and are not expected to sum to 100%. Low-Yield Rate measures the proportion of executed turns that acquire no newly matched atomic unit. Bold values indicate the best result among models in the main comparison.

patient-simulator backbone while fixing the doctor policy and evaluator.

4.4 Compared Baselines

We compare MEDCIA with general-purpose LLMs, medical-domain models, and task-related adapted baselines under the same test split, turn budget, simulator, evaluator, and task instruction. General-purpose baselines include GPT-5.4², Qwen3.6-Max-Preview³, DeepSeek-V4-Flash⁴, and Baichuan-M3-Plus⁵; medical-domain baselines include MedGemma-4B-it and MedGemma-27B-text-it (Sellersgren et al., 2025). Task-related adapted baselines include DoctorAgent-RL (Feng et al., 2025) and Note2Chat-MT (Zhou et al., 2026). API identifiers, access routes, decoding parameters, and adaptation details appear in Appendix D.5 and Appendix C.2.

4.5 Main Results

Table 1 reports the main comparison under the $K = 15$ question budget. MEDCIA achieves

²<https://openai.com/index/introducing-gpt-5-4/>

³<https://qwen.ai/blog?id=qwen3.6-max-preview>

⁴<https://api-docs.deepseek.com/news/news260424>

⁵<https://platform.baichuan-ai.com/docs/api>

the strongest performance on final importance-weighted coverage, early acquisition, and coverage speed, showing that general language ability or medical-domain pretraining alone is insufficient for cold-start longitudinal information acquisition. Its improvement on Cov@15 further indicates broader profile recovery rather than gains from only a small set of high-weight units.

Paired two-sided tests over the 200 shared test cases further confirm the improvements over GPT-5.4 on all four core information-acquisition metrics: wCov@15 (+10.4, 95% CI [8.4, 12.4]), wAUC@15 (+5.3, [3.9, 6.7]), wEarly@5 (+6.7, [5.1, 8.4]), and Cov@15 (+7.8, [6.1, 9.5]), with all Holm-adjusted $p < .001$.

Although MEDCIA does not yield the lowest HallucinationRate, its higher coverage and low ContradictionRate suggest a coverage–reliability trade-off. Many unsupported positives are elicitation-induced: overly specific or leading doctor questions may invite confirmations that are not supported by the hidden profile. This reflects a practical intake risk, since leading questions can bias patient responses, especially among older adults or patients with limited recall.

Blinded clinician evaluation. To complement the automatic evaluation, two physicians independently and blindly compared MEDCIA, GPT-5.4, SFT-only, and DeepSeek-V4-Flash on 20

anonymized cases, ranking information coverage and clinical usefulness and annotating unsupported statements and contradictions. MEDCIA achieved the highest Coverage and Usefulness Borda scores (2.425 and 2.300) and was ranked above GPT-5.4 in 77.5% of coverage comparisons (95% CI [62.5, 90.0]) and 62.5% of usefulness comparisons ([45.0, 80.0]). No unsupported statements were annotated for MEDCIA, and its mean contradiction-severity score was 0.050. Inter-rater Kendall’s τ was 0.700 for coverage and 0.517 for usefulness.

As shown in Figure 3, MEDCIA exhibits a more clinically oriented acquisition pattern than general-purpose, medical-specialized, and training-stage baselines. It achieves strong coverage on patient-reportable categories such as demographics, past medical history, symptoms, medications, allergies, and family history. These categories capture distributed longitudinal evidence usually needed in cold-start intake before downstream diagnosis or decision-making. Compared with models that concentrate on a narrower subset of categories, MEDCIA allocates turns more effectively across clinically meaningful domains, suggesting that utility-aware policy refinement improves not only the amount of recovered information but also the structure of acquired evidence.

The measurement category illustrates this policy behavior. In cold-start intake, laboratory, imaging, and detailed test measurements are useful when clinically motivated, but should not dominate early information gathering at the expense of patient-reportable history. MEDCIA achieves non-trivial measurement coverage while maintaining stronger coverage on core history-taking domains, indicating that it does not simply avoid measurement-related evidence. Instead, it balances necessary measurement probing with broader patient-background acquisition under the fixed turn budget. Together with the weighted coverage curves in Figure 4, these results suggest that MEDCIA learns a more efficient cold-start acquisition policy: it rapidly accumulates high-importance clinical evidence while preserving broad coverage across patient-reportable intake categories.

Qualitative case comparison. Table 8 presents a condensed cold-start case with end-stage renal disease and prior TAVR.

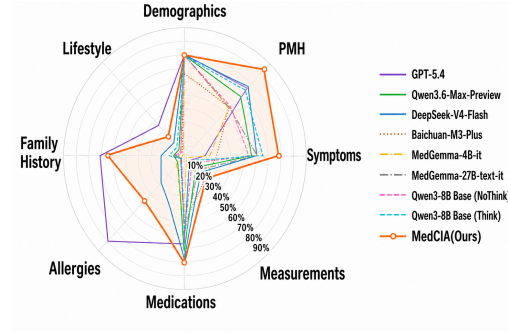


Figure 3: Category-wise clinical evidence acquisition under the cold-start intake setting. Each axis reports the macro-averaged coverage rate for one intake category, illustrating how different models allocate limited turns across patient-reportable clinical domains.

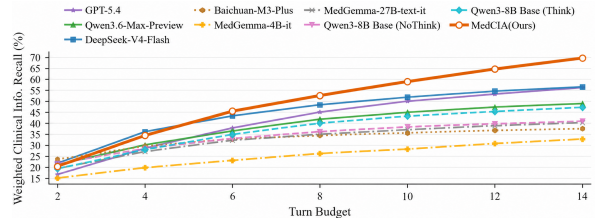


Figure 4: Weighted coverage curves $wCov_p(t)$ over dialogue turns under the $K = 15$ budget.

4.6 Ablation Studies

Table 2 compares Qwen3-8B Base in non-thinking and thinking modes, the SFT-only checkpoint, a matched PPO baseline, and full MEDCIA. PPO achieves 64.15 $wCov@15$, compared with 56.7 for SFT-only and 69.7 for full MEDCIA. Paired two-sided tests further show that MEDCIA significantly improves over SFT-only on all five reported metrics after Holm correction (all adjusted $p < .001$). Inference-mode and checkpoint details are provided in Appendix D.6.

5 Discussion

MEDCIA highlights the distinction between static medical knowledge, general multi-turn inquiry, and cold-start patient information acquisition. Static benchmarks test whether a model knows medical facts, and dynamic dialogue tests whether it can interact over turns. Cold-start intake is stricter: the agent must decide which patient-reportable evidence to elicit first under a limited question budget. Effective performance therefore requires not only clinical relevance recognition, but also efficient sequential acquisition.

Our results suggest that clinical intake should be treated as cost-sensitive evidence elicitation

Table 2: Ablation study of MEDCIA across training stages. All models are evaluated under the same test set, patient simulator, evaluator, maximum turn budget, and early-stopping rule.

Model Variant	wCov@15	wAUC@15	wEarly@5	Low-Yield Rate ↓	Precision
Qwen3-8B Base (NoThink)	41.5	33.8	31.0	73.6	72.1
Qwen3-8B Base (Think)	48.6	37.1	32.0	63.9	93.0
SFT-only	56.7	40.9	35.2	60.5	84.1
SFT + PPO	64.2	45.2	38.7	47.0	88.4
MEDCIA	69.7	47.1	40.3	45.1	89.8

rather than a byproduct of diagnosis-oriented dialogue. Under the fixed $K = 15$ budget, MEDCIA achieves stronger final weighted coverage, early weighted coverage, and coverage accumulation speed. These gains indicate that profile-grounded trajectory reconstruction and utility-aware policy refinement help recover clinically meaningful evidence across the full dialogue budget, while balancing broad screening of longitudinal background with targeted follow-up on risk-relevant cues.

The weighted-coverage formulation separates useful evidence acquisition from plausible consultation language. By decomposing hidden profiles into atomic, verifiable units, MEDINTAKE measures whether a dialogue recovers concrete patient facts. Factual reliability metrics further distinguish supported evidence from unsupported or contradictory statements, evaluating both acquisition utility and profile-grounded reliability.

6 Conclusion

We study clinical cold-start information acquisition, where clinically relevant evidence must be recovered through limited-turn interaction from little prior patient information. We formulate it as budget-constrained sequential decision-making and propose MEDCIA, a two-stage framework combining trajectory-reconstruction SFT with utility-aware policy refinement through adapted ATPO. Grounded in structured patient profiles, MEDCIA learns to prioritize patient-reportable evidence, balance broad screening with targeted follow-up, and use turns efficiently.

Under a unified protocol, MEDCIA achieves the best final importance-weighted coverage, early acquisition efficiency, and coverage accumulation speed among compared systems, with fact-level reliability diagnostics. Training-stage analysis shows complementary benefits from supervised trajectory learning and policy refinement. More broadly, the setting reflects older-adult and chronic-care intake scenarios where dense longitudinal backgrounds

must be recovered under limited interaction capacity. Future work will extend the framework to semi-open intake, cross-template generalization, reliable uncertainty expression, and human-in-the-loop verification for responsible clinical deployment.

Limitations

This work has several limitations. First, our evaluation relies on a profile-anchored closed-world simulator. While this design enables controlled comparison with progressive disclosure, incomplete recall, uncertainty expressions, and diverse response styles, it cannot fully capture the variability, emotional nuance, misunderstanding, or imperfect recall present in real clinical encounters. Results on the current benchmark should therefore be interpreted as evidence of improved controlled information acquisition, rather than direct evidence of real-world clinical effectiveness.

Second, weighted coverage depends on the pre-defined profile schema, atomic-unit construction, and evidence-based importance weights. We include a small clinician face-validity check for atomic-unit definitions and importance bins in Appendix C.1, but this check is limited in scale and should not be interpreted as full specialty-specific calibration. Future work should expand this validation to larger clinician panels, specialty-specific weighting schemes, and outcome-linked importance calibration.

Third, although the fixed evaluator enables consistent and model-blinded comparison across systems, our blinded clinician evaluation covers 20 cases and does not provide full clinician-adjudicated validation of evaluator extraction, gold-unit matching, hallucination labeling, or contradiction labeling. This limits the construct validity of absolute metric values, although relative comparisons are controlled by using the same blinded evaluator, prompts, normalization rules, hidden profiles, gold-unit inventory, and weighting file across all systems. Because the fixed evaluator and

the trainable backbone both belong to the Qwen model family, evaluator-family bias is additionally assessed through a non-Qwen evaluator-backbone sensitivity analysis in Appendix D.2, where patient responses are held fixed and only the evaluator backbone is changed. We further include a separate patient-simulator sensitivity check that fixes the Qwen evaluator and the same Qwen3-8B Think doctor while varying the patient-simulator backbone; the resulting trends remain broadly similar across 200 commonly completed simulator-sensitivity cases.

Fourth, hallucination and contradiction metrics in our closed-world setting mainly measure profile-grounded documentation controllability. They do not fully capture broader open-world clinical safety risks, such as inappropriate triage suggestions, missed emergency cues, over-reliance by users, or failures caused by ambiguous patient communication.

Fifth, this study focuses on pre-diagnostic history acquisition rather than diagnosis, treatment planning, or patient outcomes. MEDCIA is intended to support structured information collection for clinician review, not to replace clinical judgment. Future work should evaluate whether improved cold-start information acquisition translates into measurable benefits for clinician workflow, documentation quality, referral preparation, and patient safety.

Ethics Statement

This work targets pre-diagnostic information acquisition and structured history collection. MEDCIA is not designed to provide final diagnoses, treatment recommendations, prescriptions, or autonomous clinical decisions. Any real-world use should remain under the oversight of qualified clinicians.

The source records are derived from PhysioNet credentialed-access MIMIC resources under the corresponding data use agreement. These records are de-identified, and we do not attempt re-identification, date recovery, or linkage to external data sources. Due to MIMIC data-use restrictions, we will not publicly release patient-specific profiles, dialogue trajectories, gold units, or processed benchmark records through unrestricted platforms. We will release non-patient-specific artifacts, including schemas, prompts, evaluation code, metric implementations, synthetic examples,

documentation, and reconstruction scripts for credentialed MIMIC users, and will explore a PhysioNet credentialed-access release for the processed benchmark.

Potential deployment requires particular caution for older adults, patients with low health literacy, patients with limited communication ability, and patients presenting with urgent or high-risk symptoms. Cues such as chest pain, severe dyspnea, altered consciousness, acute neurological deficits, suicidal ideation, or severe allergic reactions should trigger risk escalation and human handoff rather than routine information completion. The system is intended to assist structured information collection for clinician review, not to replace clinician judgment.

References

- Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li-Jia Li, Ramesh Jain, and Amir M. Rahmani. 2024. [Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai](#). *npj Digital Medicine*, 7:82.
- American Diabetes Association Professional Practice Committee for Diabetes. 2026. [Standards of care in diabetes—2026](#). *Diabetes Care*, 49(Supplement 1).
- American Geriatrics Society Beers Criteria® Update Expert Panel. 2023. [American geriatrics society 2023 updated AGS beers criteria® for potentially inappropriate medication use in older adults](#). *Journal of the American Geriatrics Society*, 71(7):2052–2081.
- Donna K. Arnett, Roger S. Blumenthal, Michelle A. Albert, Andrew B. Buroker, Zachary D. Goldberger, Ellen J. Hahn, Cheryl Dennison Himmelfarb, Amit Khera, Donald Lloyd-Jones, J. William McEvoy, Erin D. Michos, Michael D. Miedema, Daniel Munoz, Sidney C. Smith, Salim S. Virani, Kim A. Williams, Joseph Yeboah, and Boback Ziaeeian. 2019. [2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: A report of the american college of cardiology/american heart association task force on clinical practice guidelines](#). *Circulation*, 140(11):e596–e646.
- 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*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.
- Joshua A. Boyce, Amal Assa’ad, A. Wesley Burks, Stacie M. Jones, Hugh A. Sampson, Robert A. Wood,

- Marshall Plaut, Susan F. Cooper, Matthew J. Fenton, S. Hasan Arshad, and 1 others. 2010. [Guidelines for the diagnosis and management of food allergy in the united states: Report of the NIAID-sponsored expert panel](#). *Journal of Allergy and Clinical Immunology*, 126(6):S1–S58.
- Fiona C. Bull, Salih S. Al-Ansari, Stuart Biddle, Katja Borodulin, Matthew P. Buman, Greet Cardon, Catherine Carty, Jean-Philippe Chaput, Sebastien Chastin, Roger Chou, and 1 others. 2020. [World health organization 2020 guidelines on physical activity and sedentary behaviour](#). *British Journal of Sports Medicine*, 54(24):1451–1462.
- Ruike Cao, Shaojie Bai, Fugen Yao, Liang Dong, Jian Xu, and Li Xiao. 2026. [ATPO: Adaptive tree policy optimization for multi-turn medical dialogue](#). In *International Conference on Learning Representations*.
- Mary E. Charlson, Peter Pompei, Kathy L. Ales, and C. Ronald MacKenzie. 1987. [A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation](#). *Journal of Chronic Diseases*, 40(5):373–383.
- Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. 2022. [Diaformer: Automatic diagnosis via symptoms sequence generation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4432–4440.
- Heather H. Cheng, Veda N. Giri, Michael Goggins, Matthew B. Yurgelun, Beth Y. Karlan, Barbara S. Norquist, Mary B. Daly, Tuya Pal, Zahraa AlHilli, Banu Arun, and 1 others. 2026. [Genetic/familial high-risk assessment: Breast, ovarian, pancreatic, and prostate, version 2.2026: Featured updates to the NCCN guidelines](#). *JNCCN Journal of the National Comprehensive Cancer Network*, 24(2):2–10.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Tri Dao. 2024. [FlashAttention-2: Faster attention with better parallelism and work partitioning](#). In *International Conference on Learning Representations*.
- Yue Fang, Yuxin Guo, Jiaran Gao, Hongxin Ding, Xinke Jiang, Weibin Liao, Yongxin Xu, Yinghao Zhu, Zhibang Yang, Liantao Ma, Junfeng Zhao, and Yasha Wang. 2026. [Toward better EHR reasoning in LLMs: Reinforcement learning with expert attention guidance](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30690–30698.
- Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. [A dataset of simulated patient-physician medical interviews with a focus on respiratory cases](#). *Scientific Data*, 9(1):313.
- 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](#). *Preprint*, arXiv:2505.19630. ArXiv:2505.19630.
- Jennifer Hepburn, Lynn Williams, and Lisa McCann. 2025. [Barriers to and facilitators of digital health technology adoption among older adults with chronic diseases: Updated systematic review](#). *JMIR Aging*, 8:e80000.
- Institute for Safe Medication Practices. 2024. [ISMP list of high-alert medications in acute care settings](#). Institute for Safe Medication Practices.
- Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2023. [DeepSpeed Ulysses: System optimizations for enabling training of extreme long sequence transformer models](#). *Preprint*, arXiv:2309.14509. ArXiv:2309.14509.
- Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger G. Mark. 2023a. [MIMIC-IV-Note: Deidentified free-text clinical notes](#). *PhysioNet*. Version 2.2.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023b. [MIMIC-IV, a freely accessible electronic health record dataset](#). *Scientific Data*, 10(1):1.
- Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. 2025. [Doctor-R1: Mastering clinical inquiry with experiential agentic reinforcement learning](#). *Preprint*, arXiv:2510.04284. ArXiv:2510.04284.
- Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. [MedDG: An entity-centric medical consultation dataset for entity-aware medical dialogue generation](#). In *Natural Language Processing and Chinese Computing*, volume 13551 of *Lecture Notes in Computer Science*, pages 447–459. Springer, Cham.
- Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Griseld Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. [MedSynth: Realistic, synthetic medical dialogue-note pairs](#). *Preprint*, arXiv:2508.01401. ArXiv:2508.01401.
- Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. 2018. [Ray: A distributed framework for emerging AI applications](#). In *Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation*, pages 561–577, Carlsbad, CA, USA. USENIX Association.

- Antonella Muraro, Margitta Worm, Cherry Alviani, Victoria Cardona, Audrey DunnGalvin, Lene Heise Garvey, Carmen Riggioni, Debra de Silva, Elizabeth Angier, Stefania Arasi, and 1 others. 2022. [EAACI guidelines: Anaphylaxis \(2021 update\)](#). *Allergy*, 77(2):357–377.
- National Institute for Health and Care Excellence. 2016. [Multimorbidity: Clinical assessment and management](#). NICE Guideline NG56. Published 21 September 2016.
- Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. [PriMock57: A dataset of primary care mock consultations](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.
- Gregory C. Pope, John Kautter, Randall P. Ellis, Arlene S. Ash, John Z. Ayanian, Lisa I. Iezzoni, Melvin J. Ingber, Jesse M. Levy, and John Robst. 2004. [Risk adjustment of medicare capitation payments using the CMS-HCC model](#). *Health Care Financing Review*, 25(4):119–141.
- Pengcheng Qiu, Chaoyi Wu, Junwei Liu, Qiaoyu Zheng, Yusheng Liao, Haowen Wang, Yun Yue, Qianrui Fan, Shuai Zhen, Jian Wang, Jinjie Gu, Yanfeng Wang, Ya Zhang, and Weidi Xie. 2025. [Evolving interactive diagnostic agents in a virtual clinical environment](#). *Preprint*, arXiv:2510.24654. ArXiv:2510.24654.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. [ZeRO: Memory optimizations toward training trillion parameter models](#). In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE Press.
- Sabbir M. Rashid and Deborah L. McGuinness. 2025. [Designing and evaluating an ensemble reasoning-based clinical decision support system](#). *Data Intelligence*, 7(1):1–39.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. [DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3505–3506. Association for Computing Machinery.
- Rui Reis, Pedro Rangel Henriques, João Ferreira-Coimbra, Eva Oliveira, and Nuno F. Rodrigues. 2025. [Using medical algorithms for task-oriented dialogue in LLM-based medical interviews](#). *Preprint*, arXiv:2510.12490. ArXiv:2510.12490.
- 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*, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cian Hughes, Charles Lau, and 1 others. 2025. [MedGemma technical report](#). *Preprint*, arXiv:2507.05201. ArXiv:2507.05201.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. [HybridFlow: A flexible and efficient RLHF framework](#). In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297. Association for Computing Machinery.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schackermann, Amy Wang, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, and 1 others. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31:943–950.
- Xinge Tao, Shuya Zhou, Kai Ding, Sairan Li, Yanzeng Li, Boyou Wu, Qirui Huang, Wangyue Chen, Muzi Shen, En Meng, Xiaowang Chen, Hong Hu, Jinchao Zhang, Shasha Han, and 1 others. 2026. [An LLM chatbot to facilitate primary-to-specialist care transitions: A randomized controlled trial](#). *Nature Medicine*, 32:934–942.
- Tao Tu, Mike Schackermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semurs, and 7 others. 2025. [Towards conversational diagnostic artificial intelligence](#). *Nature*, 642(8067):442–450.
- US Preventive Services Task Force. 2018. [Screening and behavioral counseling interventions to reduce unhealthy alcohol use in adolescents and adults: US preventive services task force recommendation statement](#). *JAMA*, 320(18):1899–1909.
- US Preventive Services Task Force. 2020. [Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults with cardiovascular risk factors: US preventive services task force recommendation statement](#). *JAMA*, 324(20):2069–2075.
- US Preventive Services Task Force. 2021. [Interventions for tobacco smoking cessation in adults, including pregnant persons: US preventive services task force recommendation statement](#). *JAMA*, 325(3):265–279.
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. [NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes](#).

- In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201, Bangkok, Thailand. Association for Computational Linguistics.
- Kathleen B. Watson, Jennifer L. Wiltz, Kunthea Nhim, Rachel B. Kaufmann, Craig W. Thomas, and Kurt J. Greenlund. 2025. [Trends in multiple chronic conditions among US adults, by life stage, behavioral risk factor surveillance system, 2013–2023](#). *Preventing Chronic Disease*, 22:240539.
- Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. [Task-oriented dialogue system for automatic diagnosis](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.
- WHO Collaborating Centre for Drug Statistics Methodology. 2026. [Guidelines for ATC classification and DDD assignment, 2026](#). WHO Collaborating Centre for Drug Statistics Methodology, Oslo.
- World Health Organization. 2015. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Fifth Edition, 2016*. World Health Organization, Geneva, Switzerland.
- Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024. [Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6796–6814, Bangkok, Thailand. Association for Computational Linguistics.
- Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. [End-to-end knowledge-routed relational dialogue system for automatic diagnosis](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388. ArXiv:2505.09388.
- Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. [ACI-BENCH: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation](#). *Scientific Data*, 10(1):586.
- 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*, pages 9241–9250, Online. Association for Computational Linguistics.
- Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. [PyTorch FSDP: Experiences on scaling fully sharded data parallel](#). *Proceedings of the VLDB Endowment*, 16(12):3848–3860.
- Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2024. [SGLang: Efficient execution of structured language model programs](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 62557–62583.
- Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. [Hierarchical reinforcement learning for automatic disease diagnosis](#). *Bioinformatics*, 38(16):3995–4001.
- Yang Zhou, Zhenting Sheng, Mingrui Tan, Yuting Song, Jun Zhou, Yu Heng Kwan, Lian Leng Low, Yang Bai, and Yong Liu. 2026. [Note2Chat: Improving LLMs for multi-turn clinical history taking using medical notes](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):35149–35157.
- Nana Zhu, Zixian Feng, Hang Wang, Xing Gao, Xinyi Wang, and Yuanxing Liu. 2025. [Clarifying user preference with maximum entropy based recommendation](#). *Data Intelligence*, 7(2):527–548.

A Dataset and Dialogue Construction Details

A.1 Patient Profile Construction and Dataset Validity

Profile sources and purpose. We construct patient profiles from credentialed-access MIMIC clinical resources, using MIMIC-IV-Note discharge summaries as the primary textual source and MIMIC-IV structured encounter information as complementary metadata (Johnson et al., 2023b,a). MIMIC-IV-Note provides de-identified free-text clinical notes, and discharge summaries contain rich information about admission reasons, hospital courses, prior history, medications, allergies, procedures, and discharge-related documentation. We use these records as realistic clinical sources for constructing structured patient backgrounds. Authoritative medical knowledge resources are used only to define the profile schema, field coverage, normalization rules, and importance-weighting constraints; they are not used to introduce patient-specific facts.

Data processing, temporal aggregation, and patient-reportable filtering. Each raw source record corresponds to one clinical encounter, and a patient may have multiple encounters. We therefore process records at both the encounter and patient levels. At the encounter level, we extract early-encounter cues that can realistically seed a cold-start dialogue, such as the initial emergency or admission-related presentation, chief complaint, and information explicitly available at or before presentation. At the patient level, we chronologically aggregate longitudinal background facts from prior or already-known encounters, including known diagnoses, chronic conditions, long-term medications, allergies, prior procedures or hospitalizations, family history, social history, and other long-term management factors.

For each target encounter, we retain as acquisition targets only facts that could plausibly be reported by the patient or were explicitly documented as known before or at presentation. Post-presentation clinical conclusions, newly obtained test results, and clinician-only interpretations are excluded from the gold acquisition targets unless they correspond to a known prior condition or patient-reportable event. This filtering reduces future-information leakage and aligns the benchmark with realistic cold-start intake.

De-identification, access, and release. Due to MIMIC data-use restrictions, we will not publicly release patient-specific profiles, dialogue trajectories, gold units, or profile-level splits derived from MIMIC through unrestricted platforms. We plan to release non-patient-specific artifacts, including schemas, prompts, evaluation code, metric implementations, synthetic examples, and reconstruction scripts for credentialed MIMIC users. We will also explore releasing the processed dialogue benchmark through an approved PhysioNet credentialed-access mechanism, subject to institutional and PhysioNet review.

Artifact terms and intended use. Our use and distribution of artifacts follow the access conditions of the underlying resources. MIMIC-derived records are used only for research-oriented profile construction, benchmark creation, training, and controlled evaluation under the corresponding PhysioNet credentialed-access data use agreement. We will release only non-patient-specific reproducibility artifacts through unrestricted platforms, including schemas, prompts, metric implementations,

evaluation scripts, configuration files, synthetic examples, and reconstruction scripts for credentialed MIMIC users. Patient-specific profiles, dialogue trajectories, gold units, profile-level splits, and trained artifacts derived from MIMIC will not be publicly distributed through unrestricted platforms. The intended use of MEDCIA is structured pre-diagnostic information collection for clinician review; it is not intended for diagnosis, treatment planning, prescription, or autonomous clinical decision-making.

Agent-assisted construction pipeline. We use an agent-assisted pipeline for offline profile construction and quality control. The pipeline is constrained to source-supported information: it does not introduce patient-specific facts beyond the underlying records and is not used during validation or test-time dialogue generation. First, a case-screening module identifies whether the source record contains sufficient patient-reportable information for cold-start intake simulation. Second, a quality-assessment module checks medical coherence, internal consistency, schema compatibility, and suitability for multi-turn dialogue construction, excluding cases with insufficient longitudinal background or poor schema alignment. Third, a profile-writing module extracts and normalizes source-supported facts into structured slots, including presenting complaint, history of present illness, past medical history, medications, allergies, prior procedures or hospitalizations, family history, social history, and long-term management factors.

Clinical and longitudinal coverage. The final corpus covers a broad range of clinical scenarios, containing 2,400 patient profiles, 12,171 diagnosis records, 3,163 unique diagnosis names, and 1,193 unique ICD-10 codes spanning 21 ICD-10 chapters ([World Health Organization, 2015](#)). The profiles cover 17 schema-level clinical categories, including cardiovascular, respiratory, gastrointestinal, renal/urinary, endocrine/metabolic, hematologic, neurologic, psychiatric, musculoskeletal, dermatologic, infectious, oncologic, immune/allergic, reproductive, ophthalmologic/ENT, general/systemic, and social/behavioral risk categories. Beyond disease coverage, each profile covers an average of 5.33 out of 7 long-term management dimensions, including multi-system comorbidities, polypharmacy, prior surgeries or hospitalizations, medication allergies, family history, social history, and lifestyle-related risk factors. This breadth suggests

that MEDINTAKE supports evaluation across diverse longitudinal clinical backgrounds rather than a narrow set of diseases or history domains. We use profile-level splitting throughout the paper, with 2,000 profiles for training, 200 for development, and 200 for testing. No patient profile appears in more than one split.

Inquiry-structure characterization. For dataset characterization, we assign each profile a dominant inquiry-structure category based on the distribution of patient information and the type of follow-up most needed for effective cold-start intake. This categorization is descriptive and is not used as a model input, training target, or evaluation signal. DEEP PROBING profiles contain high-risk or high-importance cues that benefit from targeted follow-up, such as severe allergy history, major comorbidities, medication-safety concerns, or red-flag symptoms. BROAD COVERAGE profiles require systematic completion of longitudinal background fields, such as chronic disease history, medication and allergy history, family history, and social or exposure-related risks. DEEP PROBING and BROAD COVERAGE profiles account for 60.1% and 39.9% of the dataset, respectively, indicating that MEDINTAKE contains both risk-focused and breadth-oriented cold-start intake scenarios.

A.2 Profile-Anchored Dialogue Corpus Construction Details

This section provides additional details on how profile-anchored dialogue trajectories are constructed for supervised learning, calibration, and corpus analysis. All construction signals that rely on hidden profiles are used only offline and are not available during validation, testing, or model inference.

Clinician-revised reference trajectories. For a small set of cases, experienced clinicians conduct profile-blind cold-start consultations with profile-anchored simulated patients under the same K -turn budget. During the initial dialogue, clinicians do not observe the hidden profile and are instructed to elicit clinically important information through focused questions. Afterward, the complete profile is revealed, and clinicians revise the trajectory by improving question wording, reordering turns, replacing low-yield questions, and adding necessary follow-ups within the same budget when appropriate. This process yields paired *Profile-Blind Trajectories* and *Clinician-Revised Reference Trajectories*,

which serve as a small high-quality reference set for quality control and trajectory calibration.

Clinicians also review the revised trajectories for clinical relevance, ordering, redundancy, and appropriateness under the limited-turn setting. For qualitative corpus analysis, reviewed questions may be annotated as DEEP PROBING or BROAD COVERAGE. These annotations are not used as model inputs, training targets, reward signals, or evaluation metrics; they are also distinct from evidence-based atomic-unit weights, which are assigned from the frozen weighting file.

Scalable simulated trajectory generation. To scale the corpus, we use profile-anchored dual-agent self-play. The doctor agent asks one focused question per turn, and the patient simulator responds under closed-world and incremental-disclosure constraints. During offline corpus construction, hidden profiles are used only to estimate the potential value of candidate questions and select high-quality supervision targets. Candidate questions are ranked by projected importance-weighted coverage gain, contextual relevance, timing appropriateness, and redundancy control. This oracle-assisted scoring is restricted to offline training-data construction and trajectory calibration; it is never available during validation, testing, or inference.

Use in training. Clinician-revised reference trajectories provide a small high-quality reference set of 80 trajectories for quality control and calibration, while large-scale profile-grounded simulated trajectories provide scalable turn-level supervision across diverse patient backgrounds and response styles. SFT instances are constructed from dialogue-history prefixes and next-question targets, and the main training configuration is reported in Appendix B.2.

A.3 Patient Simulator Validity Check

We audit the profile-anchored patient simulator on a random sample of 100 patient profiles and their generated dialogues. Each dialogue is rated by two annotators on a 1–5 Likert scale for naturalness, answerability, progressive disclosure, and profile faithfulness. In addition, we audit patient responses for unsupported assertions, contradictions against the hidden profile, and excessive information release beyond the per-turn disclosure constraint. This audit is used only to assess simulator quality and is not used to train or rank evaluated models.

The simulator achieves an average naturalness score of 4.73/5.00 and an average profile-faithfulness score of 4.95/5.00. Across 5,333 traceable profile-supported information units, only 6.8% are disclosed in the first turn, and the average first-appearance turn is 6.97, indicating gradual information release rather than upfront profile leakage. At the response level, 97.3% of audited patient responses contain no detected fabricated or contradictory patient-side facts. These results support the use of the simulator as a controlled profile-grounded evaluation environment.

A.4 Intrinsic Quality Check of Reconstructed Dialogues

To assess the intrinsic quality of the reconstructed dialogue data, we conduct a small-scale paired comparison between trajectories generated by our reconstruction pipeline and those generated by Note2Chat-MT on the same clinical notes. For each paired case, we set the evaluation budget to the smaller number of available doctor questions between the two dialogues, with a maximum cut-off of $K = 15$ turns. This matched-budget setting avoids favoring either source due to unequal dialogue length.

We report both Sentence F1@matched- K , which is compatible with the Note2Chat-style sentence-level evaluation, and profile-level Weighted Atomic Coverage@matched- K , which is computed post hoc from the corresponding patient profiles. Across 20 paired cases, our reconstructed trajectories achieve higher Sentence F1@matched- K than Note2Chat-MT (0.793 vs. 0.702) and higher profile-level Weighted Atomic Coverage@matched- K (0.667 vs. 0.547). These results suggest that the reconstructed trajectories contain more profile-aligned information under the same matched question budget. This analysis is intended only as a small-scale intrinsic data-quality sanity check, rather than a primary benchmark comparison, because it is conducted on a limited paired subset and does not evaluate trained models under the full shared simulator and evaluator protocol. The main model-level comparison is reported separately in Section 4.

We further evaluate dialogue quality using a rubric-based Clinical Dialogue Reconstruction Quality Evaluation (CDRQE) protocol across eight dimensions: clinical relevance, question ordering, inter-turn coherence, medical linkage between questions, information gain and non-redundancy,

question clarity and answerability, key risk identification, and dialogue naturalness and patient experience. Our reconstructed trajectories achieve higher overall inquiry quality than Note2Chat-MT (87.20 vs. 78.80 on a 0–100 scale), with stronger medical linkage between questions (17.70/20 vs. 15.30/20). These results provide additional evidence that the reconstructed trajectories better align with cold-start active information acquisition.

B Method and Optimization Details

B.1 Trajectory Reconstruction and SFT Details

Trajectory reconstruction. Given an ordered dialogue $D = ((Q_1, A_1), \dots, (Q_T, A_T))$ and its hidden profile p , we construct a turn-level trajectory entry for each doctor question:

$$\tau_t = (h_{t-1}, Q_t, A_t, \Delta_t, g_p(t), d_t) \quad (9)$$

where h_{t-1} is the dialogue history before turn t , Q_t is the next doctor question, A_t is the corresponding patient response, Δ_t denotes the newly covered gold units, and d_t denotes implementation-side diagnostic attributes used for offline checking, such as format validity, boundary verification, and question effectiveness. The importance-weighted coverage gain is defined as

$$g_p(t) = \frac{1}{Z_p} \sum_{u \in \Delta_t} w_p(u) \quad (10)$$

These offline fields are used to organize reconstructed trajectories, analyze supervision quality, and support reward-side checking during policy refinement. They are not exposed to the model as input features. In the SFT phase, the model is supervised to predict only the next doctor question Q_t from the preceding dialogue history h_{t-1} . The target output follows a single-field structured format:

```
<DoctorQuestion>
  QUESTION_TEXT
</DoctorQuestion>
```

Let $y_{p,t}$ denote the token sequence corresponding to the target doctor question for case p at turn t . The SFT objective maximizes the conditional likelihood over all turn-level training entries $\mathcal{Z}$:

$$\max_{\theta} \sum_{(p,t) \in \mathcal{Z}} \sum_{j=1}^{|y_{p,t}|} \log p_{\theta}(y_{p,t}^{(j)} \mid h_{p,t-1}, y_{p,t}^{(<j)}) \quad (11)$$

Equivalently, the minimized negative log-likelihood is

$$\mathcal{L}_{\text{SFT}} = - \sum_{(p,t) \in \mathcal{Z}} \sum_{j=1}^{|y_{p,t}|} \log p_{\theta}(y_{p,t}^{(j)} | h_{p,t-1}, y_{p,t}^{(<j)}) \quad (12)$$

The coverage state $\mathcal{C}_t$, newly covered units Δ_t , weighted gain $g_p(t)$, and diagnostic attributes d_t are offline variables used for trajectory reconstruction, quality control, reward computation, and evaluation. They are not included in the model input or target output. During validation and testing, each evaluated model receives only the unified system prompt and dialogue history and outputs a single `<DoctorQuestion>` field; evaluator-side states are used solely for scoring and the shared stopping rule.

B.2 Training Hyperparameters

Table 3 summarizes the main SFT and ATPO training settings used for MEDCIA.

Compute environment and budget. The trainable doctor policy is initialized from Qwen3-8B, an approximately 8B-parameter backbone. All experiments were conducted in a Python 3.10 environment on a single H100 node equipped with four NVIDIA H100 GPUs (80 GB each). Each task was allocated 16 CPU cores. We used PyTorch with NCCL 2.21.5 and FlashAttention-2 (Dao, 2024), and built the training pipeline on verl/HybridFlow (Sheng et al., 2025) and Ray (Moritz et al., 2018). For distributed execution, we adopted Actor FSDP (Zhao et al., 2023) with Ulysses sequence parallelism (Jacobs et al., 2023) set to 4, and used SGLang (Zheng et al., 2024) for rollout generation with `tensor_model_parallel_size=2` and 16 agent workers. For ATPO, we set $\alpha = 0.3$ and $\tau = 1.5$. The full training pipeline required approximately 32 GPU-hours, including trajectory-reconstruction supervised fine-tuning and utility-aware policy refinement. Evaluation on the 200-case test split and sensitivity analyses required approximately 20 additional GPU-hours. For closed-source or API-only baselines, exact parameter counts and serving infrastructure are not publicly available, so we report provider-facing model identifiers and keep the evaluation protocol fixed across systems.

B.3 Reinforcement Learning Details

B.3.1 Reward and Scorer Implementation

The second training stage performs utility-aware policy refinement by adapting Adaptive Tree Policy Optimization (ATPO) (Cao et al., 2026) to cold-start clinical information acquisition. We retain ATPO’s uncertainty-aware tree expansion and value-traceback structure, but replace its original task reward with an information-acquisition reward tailored to profile-grounded clinical intake. The implemented reward uses turn-level format compliance, boundary verification, question-effectiveness scoring, and terminal weighted recall, rather than diagnostic answer correctness.

Scorer components. During policy refinement, a fixed profile-grounded scorer computes reward signals from the accumulated dialogue. It reuses the offline extractor, atomic-unit matcher, normalization rules, and coverage computation used in the evaluation pipeline. These components are fixed before training and are not updated with the policy.

The current implementation uses a simplified scoring configuration aligned with the policy-refinement code. At the turn level, the scorer combines format, verification, and effectiveness scores. At the trajectory level, it applies a terminal weighted-recall reward under direct matching within the fixed rollout budget. Detailed coefficients are reported in Table 4.

Turn-level reward. For each generated doctor turn, the implemented scorer assigns turn-level scores for format compliance, boundary verification, and question effectiveness. Let Fmt_k , Ver_k , and Eff_k denote the corresponding binary or normalized scorer outputs at turn k . The implemented turn-level reward is

$$r_k = s_{\text{fmt}} \text{Fmt}_k + s_{\text{ver}} \text{Ver}_k + s_{\text{eff}} \text{Eff}_k \quad (13)$$

where $s_{\text{fmt}} = 0.1$, $s_{\text{ver}} = 0.4$, and $s_{\text{eff}} = 0.5$. The format score rewards compliance with the required `<DoctorQuestion>` output format, the verification score checks whether the generated turn satisfies the patient-reportable intake boundary, and the effectiveness score rewards questions that are useful for progressing the cold-start information-acquisition dialogue.

Trajectory-level reward. At the end of each rollout branch, the implementation assigns a terminal reward based on weighted recall under direct

Component	Setting	Description
Supervised fine-tuning		
Backbone	Qwen3-8B	Backbone checkpoint used to initialize the doctor policy.
Training stage	Trajectory-reconstruction SFT	Supervised learning from dialogue histories to next-question targets.
Training data	2,000 training profiles	Profile-level split; development and test sets contain 200 profiles each.
Reference trajectories	80	Clinician-revised reference trajectories used for quality control and calibration.
Fine-tuning type	Full fine-tuning	All model parameters are updated; no LoRA adapter is used.
Distributed training	DeepSpeed ZeRO-3 (Rajbhandari et al., 2020; Rasley et al., 2020)	Distributed optimization for full-parameter fine-tuning.
Max sequence length	8,192	Maximum input-output sequence length.
Development split	200 profiles	Profile-level development set; no patient profile overlaps with training or test cases.
Batch size	4 per device	Gradient accumulation steps = 2.
Learning rate	2.0×10^{-5}	SFT learning rate.
Epochs	5	Number of SFT training epochs.
Scheduler / warmup	Cosine / 0.1	Cosine learning-rate schedule with warmup ratio 0.1.
Precision	bfloat16	BF16 training.
Attention	FlashAttention-2	Efficient attention implementation.
Utility-aware policy refinement with adapted ATPO		
Initial policy	SFT checkpoint	Policy initialization for the second training stage.
Reference policy	SFT checkpoint	Frozen reference policy for KL regularization.
Critic	SFT backbone + value head	Value model initialized from the SFT backbone.
Candidate branches N	4	Number of candidate doctor questions sampled at each frontier node.
Tree budget	128	Maximum retained leaf-node budget per profile-level tree.
Maximum turns K	15	Maximum doctor-question turns per rollout.
Discount factor γ	1.0	Discount factor used in value traceback.
Uncertainty mixture α_U	0.3	Mixture coefficient for uncertainty-aware expansion.
Uncertainty threshold τ_U	1.5	Threshold for triggering full expansion.
KL weight β_{KL}	0.01	KL penalty against the frozen reference policy.
Policy learning rate	1.0×10^{-6}	Learning rate for policy updates.
Critic learning rate	1.0×10^{-5}	Learning rate for critic updates.

Table 3: Training hyperparameters for MEDCIA. The first stage performs trajectory-reconstruction supervised fine-tuning, and the second stage performs utility-aware policy refinement implemented with adapted ATPO.

matching:

$$R_T = \beta_g \text{WRecall}_p(D_{1:T}) \quad (14)$$

where $\text{WRecall}_p(D_{1:T})$ denotes the weighted recall of gold atomic units recovered by the dialogue under the `direct_match` reward mode, $\beta_g = 3.0$, and $T_{\max} = 15$ is the fixed rollout budget. The reward-side recall or judge model is fixed to `qwen3.6-flash` with deterministic decoding.

B.3.2 Adaptive Tree Policy Optimization

We initialize the trainable policy π_θ from the SFT checkpoint and use a frozen SFT copy as the reference policy π_{ref} for KL regularization. The

critic V_ψ is initialized from the same backbone with a linear value head. Following the hierarchical dialogue formulation in ATPO (Cao et al., 2026), we treat each doctor turn as a high-level macro-action and each generated token as a low-level action, while redefining the reward around information acquisition rather than diagnostic answer correctness.

At a frontier dialogue state x_k , the policy samples N candidate doctor macro-actions $\{y_k^i\}_{i=1}^N$. Each candidate is executed in the closed-world patient simulator, producing a successor state x_{k+1}^i and a turn-level reward $r(x_k, y_k^i)$. The one-step lookahead value is

$$\hat{Q}(x_k, y_k^i) = r(x_k, y_k^i) + \gamma V_\psi(x_{k+1}^i) \quad (15)$$

Component	Value / implementation	Role
Turn-level reward components		
Format score	0.1	Reward for producing an output that follows the required question format.
Verify score	0.4	Reward assigned when the generated question passes the validity or boundary-verification check.
Effective score	0.5	Reward for generating an effective patient-reportable question under the intake protocol.
Trajectory-level reward and rollout control		
Correctness / recall score	3.0	Weight for terminal weighted-recall or coverage-based trajectory reward.
β_g		
Maximum rollout turns	15	Fixed rollout budget aligned with the $K = 15$ evaluation protocol.
Recall mode	direct_match	Direct matching mode used by the reward function for recall computation.
Scorer configuration		
Recall / judge model	qwen3.6-flash	Fixed model used for recall or judge-based reward components.
Judge decoding	Temperature = 0	Deterministic decoding for reward-side judge calls.
External LLM endpoint	Omitted	Local or private service endpoint is omitted from the paper; the released code will specify the configurable endpoint interface.
Inactive components in the current implementation		
Embedding repetition detection	Not used	No frozen sentence encoder ϕ or cosine threshold τ_{rep} is used in this implementation.
Explicit information-gain threshold	Not used	No separate ε_{ig} -based low-information-gain penalty is used.
Drift / flow scoring	Not used	No explicit trajectory-level drift penalty or flow reward is used in this implementation.

Table 4: Reward configuration used in the current MEDCIA policy-refinement implementation. Turn-level scoring uses format, verification, and effectiveness components, while trajectory-level scoring uses weighted recall under direct matching. Embedding-based repetition detection, explicit information-gain thresholding, drift scoring, and flow scoring are not active in this implementation.

We use two uncertainty signals for adaptive expansion. The first measures Bellman inconsistency between the critic value and the average lookahead value:

$$U_1(x_k) = \left| V_\psi(x_k) - \frac{1}{N} \sum_{i=1}^N \hat{Q}(x_k, y_k^i) \right| \quad (16)$$

The second measures action-value dispersion among candidate macro-actions:

$$U_2(x_k) = \text{Var}_{i \in [N]} [\hat{Q}(x_k, y_k^i)] \quad (17)$$

They are combined as

$$U(x_k) = \alpha_U U_1(x_k) + (1 - \alpha_U) U_2(x_k) \quad (18)$$

If $U(x_k) > \tau_U$, all N branches are retained. Otherwise, the node is pruned by retaining one sampled branch, with a small full-expansion bypass probability $p_{\text{full}} = 0.1$ to preserve exploration.

After tree expansion, target values are computed by backward traceback. For a terminal leaf state x_T , we set

$$\hat{V}(x_T) = R_T, \quad (19)$$

where R_T is the trajectory-level terminal reward defined in Eq. 14. For a non-terminal state x_k , the traceback target is

$$\hat{V}(x_k) = \frac{1}{B(x_k)} \sum_{i=1}^{B(x_k)} [r(x_k, y_k^i) + \gamma \hat{V}(x_{k+1}^i)] \quad (20)$$

where $B(x_k)$ is the number of retained child branches. The macro-action advantage is estimated as

$$\hat{A}(x_k, y_k^i) = r(x_k, y_k^i) + \gamma V_\psi(x_{k+1}^i) - V_\psi(x_k) \quad (21)$$

The expanded tree is decomposed into root-to-leaf trajectories. The policy is updated with a

clipped PPO-style objective over generated action tokens, assigning the same macro-action advantage to all tokens in the corresponding doctor turn:

$$J_{\text{ATPO}}(\theta) = \mathbb{E} \left[\frac{1}{M} \sum_{j=1}^M \frac{1}{K^{(j)}} \sum_{k=1}^{K^{(j)}} \mathcal{L}_{j,k} - \beta_{\text{KL}} D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right] \quad (22)$$

Here, M is the number of root-to-leaf trajectories and $K^{(j)}$ is the number of turns in trajectory j . For compact notation, define

$$A_{j,k} = \hat{A}(x_k^{(j)}, y_k^{(j)}), \quad C_{j,k} = C(x_k^{(j)}), \quad L_{j,k} = L(y_k^{(j)}). \quad (23)$$

For compact notation, define the clipped ratio and the token-level PPO term as

$$\bar{\rho}_{j,k,\ell} = \text{clip}(\rho_{j,k,\ell}, 1 - \epsilon, 1 + \epsilon), \quad \ell_{j,k,\ell} = \min\{\rho_{j,k,\ell} A_{j,k}, \bar{\rho}_{j,k,\ell} A_{j,k}\} \quad (24)$$

The local clipped objective for the k -th macro-action on trajectory j is then

$$\mathcal{L}_{j,k} = \frac{1}{C_{j,k} L_{j,k}} \sum_{\ell=1}^{L_{j,k}} \ell_{j,k,\ell} \quad (25)$$

The visit count $C_{j,k}$ down-weights frequently visited states, while $L_{j,k}$ normalizes the objective by the token length of the doctor macro-action.

For the token-level importance ratio, let

$$c_{j,k,\ell} = (x_k^{(j)}, y_{k,<\ell}^{(j)})$$

denote the token-generation context. The ratio is computed against the behavior policy used to generate the rollout:

$$\rho_{j,k,\ell} = \frac{\pi_{\theta}(y_{k,\ell}^{(j)} \mid c_{j,k,\ell})}{\pi_{\theta_{\text{old}}}(y_{k,\ell}^{(j)} \mid c_{j,k,\ell})} \quad (26)$$

The critic is trained to predict the traceback target $\hat{V}(x_k)$ using mean squared error.

Invalid-output handling. We apply deterministic parsing before sending a generated doctor turn to the patient simulator. If the output contains a single recoverable patient-reportable question but does not exactly follow the required `<DoctorQuestion>` format, the recovered question is used to continue the rollout while

the format score reflects the violation. If the output cannot be parsed into a single patient-reportable question, contains multiple distinct questions, or violates the intake boundary, the turn receives no effectiveness credit and does not contribute new recall credit for that turn. The rollout then continues with the invalid turn recorded in the dialogue state.

C Additional Experiment Details

C.1 Evidence-Based Importance Weighting

Importance weights are assigned offline rather than estimated by the evaluated models. Each gold atomic unit u in patient profile p is linked to a fixed importance weight $w_p(u)$ from a frozen evidence-based weighting file. We map weights into four ordered levels: *critical*, *high*, *medium*, and *low*, corresponding to 1.0, 0.7, 0.4, and 0.1, respectively. The same weighting file is used for all training-stage analysis and model evaluation.

The weighting file is constructed from clinical risk, medication-safety, allergy, family-risk, and behavioral-risk references. These include the Charlson Comorbidity Index (Charlson et al., 1987), the CMS-HCC risk adjustment model (Pope et al., 2004), ADA and NICE guidance for chronic disease and multimorbidity management (American Diabetes Association Professional Practice Committee for Diabetes, 2026; National Institute for Health and Care Excellence, 2016), the WHO ATC/DDD classification system and the ISMP High-Alert Medication List (WHO Collaborating Centre for Drug Statistics Methodology, 2026; Institute for Safe Medication Practices, 2024), the AGS Beers Criteria (American Geriatrics Society Beers Criteria® Update Expert Panel, 2023), NI-AID and EAACI guidance for allergic reactions and anaphylaxis (Boyce et al., 2010; Muraro et al., 2022), NCCN and ACC/AHA guidance for familial and cardiovascular risk assessment (Cheng et al., 2026; Arnett et al., 2019), and USPSTF/WHO recommendations for behavioral risk factors (US Preventive Services Task Force, 2020, 2021, 2018; Bull et al., 2020).

For information units that cannot be directly matched to the evidence-based tables, we use an explicitly documented category-level design prior as a fallback weight. Each gold unit records a `weight_source_ref` field linking the assigned weight to the corresponding source table, entry, and version when available, or to the fallback category prior otherwise. Because the weighting

file is fixed before evaluation and shared across all systems, weighted coverage reflects the same predefined clinical-importance scheme for every model rather than model-specific or evaluator-specific preference. The current weighting scheme is therefore designed for transparent and reproducible benchmark evaluation, but it should not be interpreted as a fully clinician-calibrated importance scale across all specialties.

Clinician face-validity check. To assess whether the weighting scheme and atomic-unit construction are clinically plausible, we conduct a small clinician face-validity check on a stratified sample of $N_{\text{unit}} = 120$ gold atomic units covering major information categories and all four importance bins. Each unit is independently reviewed by $N_{\text{reviewer}} = 2$ clinically trained reviewers for atomic-unit clarity, patient-reportability, and appropriateness of the assigned importance bin. The acceptable rate for atomic-unit definitions is 96.7%, the exact agreement with the predefined importance bin across reviewer-unit judgments is 92.9%, and the exact-or-adjacent-bin agreement is 100.0%. The bin-level inter-reviewer agreement is substantial, with Cohen’s $\kappa = 0.809$ (Cohen, 1960). These results support the clinical plausibility of our atomic-unit definitions and importance bins, while larger specialty-specific calibration remains an important direction for future work.

C.2 Baseline Descriptions and Inference Settings

Compared systems and access. For closed-source or API-accessed models, we use the provider-reported model identifiers available at the time of evaluation, rather than relying on architectural details that are not publicly disclosed. The general-purpose LLM baselines include GPT-5.4 with provider identifier `gpt-5.4` through the OpenAI API, Qwen3.6-Max-Preview with `qwen3.6-max-preview` through the Qwen API, DeepSeek-V4-Flash with `deepseek-v4-flash` through the DeepSeek API, and Baichuan-M3-Plus with `baichuan-m3-plus` through the Baichuan API. API-accessed baselines were evaluated on 2026-05-19 using the available provider interfaces.

Medical-domain baselines. We include MedGemma-4B-it and MedGemma-27B-text-it (Sellergren et al., 2025) to examine whether medical-domain pretraining and increased model

scale directly translate into more efficient cold-start information acquisition. These models are evaluated with the same task instruction, turn budget, patient simulator, evaluator, and metric computation procedure as the general-purpose baselines.

Unified inference protocol. All main-comparison systems are evaluated on the same test split under the same cold-start intake protocol. Unless a provider interface imposes different constraints, we use deterministic decoding with temperature = 0 and a maximum generation length of 1,024 tokens. Reasoning-mode settings are explicitly controlled when supported by the provider interface: GPT-5.4 is evaluated with high reasoning effort, DeepSeek-V4-Flash is evaluated with high reasoning enabled, and Qwen3.6-Max-Preview and Baichuan-M3-Plus use provider-default inference settings. All systems share the same prompt template, maximum turn budget, closed-world patient simulator, early-stopping rule, evaluator, and metric computation code. If a context-window limit is reached, the most recent eight turns are retained under the same truncation rule across affected systems.

D Evaluation Protocol and Supplementary Analyses

D.1 Full Metric Definitions

Evaluation timeline and early stopping. Each model interacts with the patient simulator for at most $K = 15$ doctor-question turns. The fixed cold-start opening question is counted as the first doctor turn within the same K -turn evaluation budget. After turn 12, the evaluator stops a dialogue if the incremental weighted gain remains no greater than $\epsilon = 0.05$ for three consecutive turns:

$$T = \min \left\{ 12 \leq t \leq K : \max_{j \in \{0,1,2\}} g_p(t-j) \leq \epsilon \right\} \quad (27)$$

with $T = K$ if the condition is never satisfied. This stopping rule is applied only by the evaluator and is not exposed to the doctor policy.

Carry-forward alignment. For dialogues terminated early at turn $T \leq K$, coverage metrics are aligned by carry-forward: for all $t > T$, we set $\mathcal{C}_t = \mathcal{C}_T$ and therefore $\text{wCov}_p(t) = \text{wCov}_p(T)$. This alignment keeps all systems comparable under the same nominal budget and ensures that wAUC@15 is well-defined.

Weighted and unweighted coverage. For patient p , let $\mathcal{U}_p$ be the gold atomic-unit set, $w_p(u)$ the fixed importance weight of unit u , $Z_p = \sum_{u \in \mathcal{U}_p} w_p(u)$ the total weight mass, and $\mathcal{C}_t$ the correctly covered gold-unit set after turn t . Final weighted coverage is

$$\text{wCov}_p(K) = \frac{1}{Z_p} \sum_{u \in \mathcal{C}_K} w_p(u) \quad (28)$$

where $K = 15$ in the main evaluation. Unweighted coverage is

$$\text{Cov@}K = \frac{|\mathcal{C}_K|}{|\mathcal{U}_p|}. \quad (29)$$

Early coverage, area under the curve, and supplementary threshold turns. Early-stage weighted coverage is defined as

$$\text{wEarly@s} = \text{wCov}_p(s) \quad s \in \{3, 5\}. \quad (30)$$

The area under the weighted coverage curve is

$$\text{wAUC@}K = \frac{1}{K} \sum_{t=1}^K \text{wCov}_p(t). \quad (31)$$

For the supplementary threshold-reaching analysis, we use a 50% coverage threshold:

$$\text{wT50}_p = \min \{t \leq K : \text{wCov}_p(t) \geq 0.5\} \quad (32)$$

If the threshold is not reached within the budget, the case is assigned $K + 1$ when computing the mean.

Factual reliability metrics. We report factual reliability as fact-level diagnostic rates with different denominators rather than as a single mutually exclusive partition. Let $\mathcal{G}_p$ denote the set of gold atomic units for patient profile p , and let $\hat{\mathcal{R}}_p$ denote the atomic units reconstructed from the dialogue. The gold-unit matching stage assigns each gold unit one of three statuses: *covered*, *missing*, or *contradicted*. Ambiguous, underspecified, or insufficiently supported cases are treated as *missing* unless the dialogue explicitly supports or contradicts the gold unit.

Unsupported reconstructed units are identified in a separate evaluator stage. These unsupported units are classified by polarity and certainty into confirmed, denied, or uncertain unsupported items. Only unsupported confirmed positive assertions are counted as hallucinations; unsupported denials and

uncertainty statements are retained for diagnostic analysis but are not counted as hallucinations.

Let $\mathcal{G}_p^{\text{cov}}$ be the set of covered gold units, and let $\hat{\mathcal{U}}_p^{\text{conf}}$ be the set of unsupported confirmed reconstructed units. We define the positive-fact denominator as

$$d_p^+ = |\mathcal{G}_p^{\text{cov}}| + |\hat{\mathcal{U}}_p^{\text{conf}}| \quad (33)$$

When $d_p^+ = 0$, both Precision and HallucinationRate are set to 0 for that case; the absence of recovered information is reflected by the coverage metrics. Otherwise,

$$\begin{aligned} \text{Precision}_p &= \frac{|\mathcal{G}_p^{\text{cov}}|}{d_p^+}, \\ \text{HallucinationRate}_p &= \frac{|\hat{\mathcal{U}}_p^{\text{conf}}|}{d_p^+} \end{aligned} \quad (34)$$

Thus, Precision and HallucinationRate are complementary under the positive-fact denominator, except in cases with no covered or unsupported confirmed facts.

ContradictionRate is computed separately over gold atomic units:

$$\text{ContradictionRate}_p = \frac{|\mathcal{G}_p^{\text{contra}}|}{|\mathcal{G}_p|}. \quad (35)$$

Because ContradictionRate uses the gold-unit denominator, whereas Precision and HallucinationRate use the positive-fact denominator, these metrics are not expected to sum to 100%. They should be interpreted as complementary diagnostics of different factual failure modes rather than as components of a single probability distribution. All reliability metrics are macro-averaged over test cases.

Low-yield behavior and supplementary quality metrics. The Low-Yield Turn Rate is defined as

$$\text{LowYieldRate} = \frac{|\{t \leq T : |\Delta_t| = 0\}|}{T}, \quad (36)$$

where T is the number of executed turns. A budget-normalized version using K as the denominator is reported as a sensitivity analysis. We also include three supplementary LLM-judged profile-level scores: LLM Completeness, LLM Faithfulness, and LLM Granularity. Each score is assigned on a 0–100 scale for each dialogue case by the fixed Stage-C evaluator and then macro-averaged over evaluated cases. These scores are used only as secondary diagnostic measures; the primary conclusions rely on profile-grounded atomic coverage and fact-level reliability metrics.

D.2 Evaluator Protocol, Quality Controls, and Sensitivity

Atomic-unit matching and normalization.

Each patient profile is decomposed into independently verifiable atomic medical units, each containing a field category, normalized fact expression, allowed synonyms, and an importance weight. During evaluation, extracted dialogue units are matched to gold units using the same predefined normalization and matching rules for all systems. These rules cover medication-name normalization, clinically equivalent expressions, polarity and negation handling, and predefined tolerance ranges for temporal, dosage, frequency, and numerical information.

Fixed extraction protocol. The profile-grounded evaluator is fixed before evaluation and applied identically to all systems. It uses deterministic decoding with temperature = 0, shared extraction prompts, rule-based normalization, cached outputs, and model-identity-blinded matching. This design keeps coverage, hallucination, contradiction, and low-yield metrics comparable under the controlled benchmark protocol.

Profile-score rubric. The evaluator receives the gold atomic units, the dialogue-reconstructed profile, and the original doctor–patient dialogue. The original dialogue is treated as the authoritative source for coverage and turn attribution, while the reconstructed profile serves as an auxiliary structured representation. The gold-unit matching stage assigns each gold unit one of three statuses: *covered*, *missing*, or *contradicted*. Ambiguous, underspecified, or insufficiently supported cases are treated as *missing*. Unsupported reconstructed-profile items and supplementary profile-level scores are handled in separate evaluator stages rather than as gold-unit statuses. After atomic matching, the evaluator assigns three supplementary profile-level scores on a 0–100 scale: Completeness, Faithfulness, and Granularity. Completeness rewards recovery of gold units and penalizes missing high-importance units. Faithfulness penalizes unsupported reconstructed items and contradicted gold units. Granularity rewards clinically specific details, such as medication name with dose or frequency, disease duration or timing, procedure timing, allergy details, and numerical clinical values. These scores are reported only as supplementary diagnostics.

Dialogue realism check. As a supplementary realism check, we compare constructed dialogues with the 80 clinician-conducted profile-blind consultation trajectories using a blinded LLM discriminator. The discriminator does not reliably separate the two sources across repeated runs, and the realism gap between clinician-conducted and constructed dialogues remains small. This result suggests that the constructed dialogues are close to clinician-style blind consultations in surface naturalness and inquiry style. We use this analysis only as auxiliary evidence for dialogue naturalness; the primary evaluation relies on the fixed profile-grounded evaluator and the atomic coverage and fact-level reliability metrics defined above.

Evaluator-family sensitivity. To examine whether the evaluation depends on a single evaluator backbone, we conduct an evaluator-family sensitivity analysis using the same main-comparison model outputs. The default evaluator is Qwen3.6-Flash, and the alternative evaluators are DeepSeek-V4-Flash with high reasoning enabled and a Llama-family evaluator. Only the evaluator backbone is changed; the model outputs, patient responses, gold profiles, gold units, importance weights, matching rules, and metric code remain fixed.

Table 5 reports model-level results under the three evaluator backbones. This analysis is used as a robustness diagnostic for evaluator-family effects, not as a separate benchmark or a substitute for clinician adjudication.

Patient-simulator sensitivity. We further examine whether the same doctor policy shows similar acquisition behavior under different patient-simulator backbones. The doctor model is fixed to Qwen3-8B Think and the evaluator is fixed to Qwen3.6-Flash. We compare locally deployed Qwen-family, DeepSeek-family, and Llama-family patient simulators under the same 200-case simulator-sensitivity set and metric computation procedure. Table 6 reports the resulting simulator-family sensitivity results. This analysis characterizes how patient-simulator choice affects absolute metric values while keeping the doctor model and evaluator fixed.

Reporting convention. Unless otherwise stated, all reported scores are averaged over repeated runs on the fixed 200-case test split. We report mean scores rather than best-run or maximum scores.

Table 5: Evaluator-family sensitivity results on the main-comparison systems. Qwen Eval. denotes the default Qwen3.6-Flash evaluator. DeepSeek Eval. denotes the DeepSeek-V4-Flash evaluator with high reasoning enabled. Llama Eval. denotes the Llama-family evaluator. Percentage-valued metrics are reported on a 0–100 scale.

Model	wCov@15 ↑			wAUC@15 ↑			Prec. ↑			Halluc. Rate ↓		
	Qwen	DeepSeek	Llama	Qwen	DeepSeek	Llama	Qwen	DeepSeek	Llama	Qwen	DeepSeek	Llama
MEDCIA	69.7	69.1	70.8	47.1	47.1	49.4	89.8	88.9	92.0	10.2	11.1	8.0
GPT-5.4	59.3	59.0	63.3	41.8	41.7	44.8	97.5	96.8	98.4	2.5	3.2	1.6
Qwen3.6-Max-Preview	50.7	50.2	56.0	38.9	38.7	43.1	73.4	69.2	87.8	26.6	30.8	12.2
DeepSeek-V4-Flash	58.3	57.8	62.6	44.7	44.5	48.0	81.9	79.7	92.1	18.1	20.3	7.8
MedGemma-4B-it	35.7	40.4	43.8	25.8	32.5	30.8	90.8	71.2	96.0	9.2	28.8	4.0
MedGemma-27B-text-it	41.2	35.4	44.0	32.9	25.6	35.0	75.5	89.1	85.8	24.5	10.9	14.2
Baichuan-M3-Plus	38.1	37.0	41.3	32.7	31.8	35.6	86.3	83.8	90.4	13.7	16.2	9.6

Table 6: Patient-simulator sensitivity under the same doctor model and fixed evaluator. Percentage-valued metrics are reported on a 0–100 scale.

Doctor Model	Patient Simulator	Cases	wCov@15	wAUC@15	wEarly@5	wT50	Cov@15	Prec.	Halluc. Rate	Low-Yield Rate
Qwen3-8B Think	Qwen patient simulator	200	48.6	37.1	31.9	12.7	44.6	93.0	7.0	62.1
Qwen3-8B Think	DeepSeek patient simulator	200	44.1	33.3	29.6	13.8	41.3	85.9	14.0	63.3
Qwen3-8B Think	Llama patient simulator	200	44.5	33.4	28.9	13.6	40.0	75.2	24.8	66.7

Each run follows the same turn budget, simulator protocol, evaluator, decoding configuration, and metric implementation. Percentage-valued metrics are reported on a 0–100 scale. For evaluator-family sensitivity analyses, we report full model-level scores under alternative evaluator backbones.

D.3 Evaluation Pipeline

Pipeline overview. The evaluation pipeline proceeds in multiple stages. First, the fixed evaluator extracts structured information units from the accumulated dialogue and records the first appearance turn of matched units. Second, the gold-unit matching stage compares the original dialogue against precomputed gold atomic units and assigns each gold unit one of three statuses: *covered*, *missing*, or *contradicted*. Ambiguous or insufficiently supported cases are treated as *missing*. Third, unsupported reconstructed items are identified separately for hallucination-related analysis. Finally, metric code computes weighted coverage curves, early-coverage metrics, threshold-reaching turns, fact-level reliability rates, low-yield behavior, and supplementary profile-level scores.

This subsection summarizes the execution order of the evaluation pipeline. Full metric definitions are provided in Appendix D.1, and evaluator normalization rules, fixed extraction protocol, profile-score rubric, and sensitivity analyses are provided in Appendix D.2. Full prompt templates and implementation details are provided in the released evaluation code rather than reproduced verbatim in

the appendix.

D.4 Supplementary Evaluation Results

Table 7 reports supplementary metrics not included in the main result table, including earlier weighted coverage, threshold-reaching turns, and profile-level reconstruction quality. These metrics are used as secondary diagnostics; the main conclusions rely on the primary coverage, efficiency, factual reliability, and low-yield metrics reported in Table 1.

D.5 Task-Related Adapted Comparisons

Within the main comparison in Section 4, we include task-related adapted baselines to contextualize MEDCIA against prior medical-dialogue paradigms. These systems were originally designed for settings different from profile-grounded cold-start information acquisition; therefore, we report them as adapted transfer baselines, while using the same 200-case test split, simulator, evaluator, turn budget, and metric code as the other main-comparison systems.

Adaptation protocol. All adapted systems are evaluated under the same cold-start intake interface as the main comparison. Each system is constrained to output one <DoctorQuestion> per turn, interacts with the same profile-anchored patient simulator, and is evaluated by the same information-centric evaluator and metric code. The adapted-comparison results reported in the main table are computed on the 200-case test split; this

Table 7: Supplementary metrics for the main comparison. The evaluated systems follow the same test split, simulator, evaluator, turn budget, and early-stopping rule as Table 1. Percentage-valued metrics are reported on a 0–100 scale.

Model	wEarly@3 ↑	wT50 ↓	LLM Completeness ↑	LLM Faithfulness ↑	LLM Granularity ↑
<i>Closed-source model</i>					
GPT-5.4	23.0	11.5	60.8	87.2	63.0
Qwen3.6-Max-Preview	26.7	12.2	51.0	81.0	54.2
DeepSeek-V4-Flash	30.7	10.0	58.9	83.4	58.5
Baichuan-M3-Plus	27.4	14.2	39.0	82.4	40.0
<i>Medical-domain open-source models</i>					
MedGemma-4B-it	24.2	14.1	42.7	79.7	44.7
MedGemma-27B-text-it	17.8	14.7	36.2	90.8	39.9
<i>Task-related adapted baselines</i>					
DoctorAgent-RL	27.0	14.2	30.8	82.8	50.8
Note2Chat-MT	27.8	14.5	32.4	88.0	39.6
<i>Profile-grounded information acquisition model</i>					
MEDCIA	27.3	8.3	70.5	84.7	59.5

Note. wEarly@3 measures importance-weighted atomic coverage after the first three doctor questions. wT50 denotes the number of turns required to reach 50% weighted coverage; unreached cases are assigned 16. LLM Completeness, LLM Faithfulness, and LLM Granularity are supplementary 0–100 profile-level scores assigned by the fixed Stage-C evaluator; they are macro-averaged over evaluated cases and are not used as primary ranking metrics.

appendix provides additional adaptation details and interpretation.

DoctorAgent-RL. DoctorAgent-RL (Feng et al., 2025) represents a reinforcement-learning-based medical consultation paradigm. We evaluate the publicly released DoctorAgent-RL model in our cold-start intake environment by enforcing the same one-question-per-turn interface and evaluating its generated dialogues with our profile-grounded information-acquisition metrics. This comparison examines how a consultation-oriented RL policy transfers to patient-reportable evidence elicitation under a fixed turn budget.

Note2Chat-MT. Note2Chat-MT (Zhou et al., 2026) represents a note-to-dialogue history-taking paradigm. Because a directly reusable public model checkpoint is not available, we reproduce its note-to-dialogue construction procedure to generate a task-related training corpus, train an adapted Note2Chat-MT model, and evaluate the resulting model under the same cold-start intake protocol. This comparison examines whether note-derived dialogue supervision transfers to proactive recovery of longitudinal patient background.

Interpretation. These results should not be interpreted as exact reproductions of the original evaluation settings of DoctorAgent-RL or Note2Chat-MT. Instead, they provide a controlled transfer comparison under MEDINTAKE’s stricter cold-start protocol, where the objective is patient-reportable information recovery rather than diagnosis, treatment recommendation, or note-to-dialogue generation alone. The main conclusions of this paper are based on the unified main-comparison results in Section 4.

D.6 Ablation Details

We report training-stage ablations to examine how performance changes across the main stages of MEDCIA. All variants are evaluated under the same test set, patient simulator, evaluator, maximum turn budget, early-stopping rule, and metric computation procedure.

Qwen3-8B Base (NoThink) evaluates the backbone model under the cold-start intake protocol without enabling the model’s explicit thinking mode when such a mode is available. Qwen3-8B Base (Think) evaluates the same backbone model with its available thinking mode enabled, allowing us to compare the effect of inference-mode configuration under the same task protocol. SFT-only eval-

uates the checkpoint after trajectory-reconstruction supervised fine-tuning, before utility-aware policy refinement, and measures the effect of supervised trajectory learning on cold-start information acquisition. MEDCIA is the full model after supervised trajectory learning and utility-aware policy refinement. It uses the same inference interface as the other evaluated systems and receives only the dialogue history and patient responses during evaluation.

E Prompt Templates

This appendix reports the exact runtime prompt templates used by the doctor agent and the profile-grounded patient simulator. We also include the core gold-unit matching prompt used in Task 1 of the evaluator pipeline, since it defines the three-way coverage status used for metric computation. Other data-construction and evaluator prompts are provided in the released code rather than reproduced verbatim in the appendix. For readability, prompts are presented as full-width boxed blocks with line wrapping.

E.1 Doctor Agent Prompt

Doctor Agent Prompt

```
You are a senior AI Doctor conducting a medical interview with a patient.

Your core strategy:
Start with ZERO knowledge about the patient.
Follow the BFS/DFS strategy: use Breadth-First Search (BFS) during initial
    scanning, and use Depth-First Search (DFS) when red flags are identified.
Connect the dots: ensure that your next question has a clear logical connection
    with the patient's previous answer.
Use implicit probing: elicit medical history by asking about "functional impact"
    or "associated symptoms" rather than asking diagnosis names too early.

[Task Requirements]
You have a total of {k_rounds} rounds of questioning.
This is a cold-start medical history acquisition task under a limited turn
    budget.
Your priority is to efficiently collect the patient's long-term health
    background.

[Advanced Questioning Guidelines]
Connecting the Dots:
Do not ask isolated questions. Ensure logical coherence between the patient's
    previous answer and your next question.

Layered Extraction:
Do not ask for all medication names at once. Follow a logical sequence:
    identify the condition -> ask about treatment/medications -> inquire about side
        effects or adherence.

Natural Transitions:
Use natural conversational bridges to make the interview feel like a real,
    professional clinical encounter.

[Constraints]
Ask only 1 focused question per turn.
Do not ask multiple questions at once.
Do not hallucinate facts that have not yet been revealed in the dialogue history.

Based on the dialogue history, reason out the next question with the highest
    information gain, prioritizing questions that can complete missing
        relationships between known facts rather than merely discovering isolated
            new facts.

[Current Round]
Round {current_round} / Total {k_rounds} rounds

[Output Format]
<DoctorQuestion>QUESTION_TEXT</DoctorQuestion>
```

E.2 Patient Simulator Prompt

The patient simulator is constrained by the hidden patient profile and prior patient-side utterances. It is designed to support progressive disclosure, uncertainty expression, and strict factual controllability.

Patient Simulator Prompt

you are a real patient being interviewed by a doctor. Answer only from the patient_profile and your earlier answers when they match it. Never mention profile, prompt, data, or system.

Speak naturally in English, 1-2 short sentences.

Rules:

- If the first question asks for chronic conditions or diagnoses, mention at most 2 major conditions, then say "and a few other things" if there are more. Do not reveal all the symptoms at once.
- Do not list more than three symptoms at one time.
- Do not invent symptoms, diagnoses, tests, dates, durations, numbers, causes, habits, side effects, or treatments.
- Say "Yes" only when the profile clearly supports it. Say "No" only when it clearly denies it. Otherwise say "I don't know."
- Answer only the thing asked this turn. Do not volunteer unasked diseases, medicines, surgeries, allergies, family history, or social history.
- Small information budget: normally give 1 fact, plus at most 1 exact detail if needed. Never give a list.
- In each round, do not provide more than 3 pieces of information, including the name of the medicine, surgical history, family history, etc.
- If a question is broad or asks several things at once, choose one directly relevant item and wait for follow-up.
- For medications, even if the doctor asks "which medications" or names several diseases, mention only one medication or one medicine purpose. Add dose or frequency only as the one extra detail. If asked for more, give only one new item.
- If information is not clearly present, answer uncertainty briefly.
- When answering, you can occasionally add interjections to make it more personified. For highly academic terms, drug names, and disease names, it is acceptable to occasionally use colloquial descriptions.

E.3 Evaluator Task-1 Matching Prompt

The following prompt is used for Task 1 of the evaluator pipeline. It matches precomputed gold atomic units against the original doctor-patient dialogue and assigns each gold unit one of three statuses: *covered*, *missing*, or *contradicted*. Data-construction prompts for decomposing gold profiles into atomic units, as well as later evaluator-side prompts for unsupported-item detection and supplementary profile scoring, are released with the evaluation code rather than reproduced here.

Gold-Unit Matching Prompt

You are a clinical information quality assessment expert.

Input Data

1. Gold Profile Information Units (gold_units):

Information units split from the patient's standard profile. Each unit has a canonical_text and may have variants. A gold unit may be a positive statement, meaning the patient HAS X, or a negative statement, meaning the patient does NOT have X, such as "No known drug allergies".

2. Original Dialogue (original_dialogue):

The raw doctor-patient dialogue. This is the ONLY authoritative source for deciding coverage and turn numbers. Every user message has a patient_turn field that records the patient's N-th utterance in the conversation. Doctor messages have patient_turn=null. The conversation's total patient turn count is given in total_patient_turns.

For each gold unit in `gold_units.units`, scan `original_dialogue` and decide its status. There are exactly 3 statuses: covered, missing, contradicted. There is no uncertain status.

You MUST output one `gold_match_results` entry per input gold unit, in the same order as `gold_units.units`, with the same `unit_id`.

Status Definitions

covered:

The patient's utterances in the dialogue support the gold unit.

Examples:

- The patient explicitly affirms the gold unit.
- The patient affirms the gold unit with hedged or tentative language.
- For a negative gold unit, the patient explicitly denies the topic.
- For medication units, mentioning the drug name with a positive sense covers the medication-identity unit. Dose, frequency, route, or formulation details require explicit or semantically implied support if they are part of the gold unit.

contradicted:

The dialogue explicitly conflicts with the gold unit. This includes positive gold plus denial, negative gold plus affirmation, or direct value conflict.

missing:

Anything that is not covered and not contradicted, including topics never raised, cases where the patient admits not knowing, or statements too vague to match.

`matched_first_turn`

`matched_first_turn` must be the `patient_turn` value of the earliest user message that supports the gold unit for covered, or contradicts it for contradicted.

- Read `original_dialogue.messages[i].patient_turn` directly.
- Do not count messages manually.
- Set `matched_first_turn` = null only for missing.
- `matched_first_turn` must be an integer in `[1, total_patient_turns]` when not null.

Output Format

Return strictly JSON with only these three top-level keys: `case_id`, `path_id`, `gold_match_results`.

Each `gold_match_results` entry must contain:

`unit_id`, `status`, `matched_dialogue_item_ids`, `matched_first_turn`, `match_type`, `match_confidence`, and `evidence`.

Do not include `unsupported_count`, `unsupported_dialogue_items`, or `profile_scores`.

Gold Profile Information Units
{`gold_units_json`}

Original Dialogue
{`original_dialogue_json`}

E.4 Qualitative Cold-Start Case Comparison

To provide a concrete example of how different inquiry policies behave under the same cold-start setting, Table 8 compares physician ordering, annotated ordering, baseline model behavior, and MEDCIA on a high-risk patient profile involving end-stage renal disease, prior TAVR, atrial fibrillation, pulmonary hypertension, chronic anemia, and dialysis-related history. This example is intended as qualitative illustration rather than a primary evaluation metric.

Table 8: Condensed cold-start case comparison for a high-risk patient with ESRD and prior TAVR.

Patient Profile	A 73-year-old male with severe aortic stenosis status post TAVR, paravalvular aortic insufficiency, hypertension, paroxysmal atrial fibrillation, heart failure with preserved ejection fraction, pulmonary hypertension, end-stage renal disease on thrice-weekly hemodialysis, chronic anemia, and prior upper gastrointestinal bleeding. He has a left subclavian vein stent and dialysis fistula, and takes amiodarone, amlodipine, torsemide, sevelamer, pantoprazole, vitamin D, senna, and a B-complex supplement.	
Task Setting	No pre-dialogue chief complaint is provided. All systems start from the same fixed cold-start opening question: “ <i>Could you briefly tell me your age, gender, and any long-term medical conditions you have?</i> ” The simulated patient follows incremental disclosure and only reveals profile-supported facts.	
	Condensed Dialogue	Consultation Evaluation
Physician Original Order	D: Age, gender, and chronic diagnoses? P: 73-year-old male; prior valve treatment and kidney failure on dialysis. D: Any long-term heart or blood-pressure medications? P: Amiodarone and amlodipine. D: How often is dialysis? P: Three times per week. D: Any drug allergy, surgery, or hospitalization history? P: No clear drug allergy; prior valve procedure and hospitalizations for kidney failure or GI bleeding. D: Any other major conditions? P: Atrial fibrillation, pulmonary hypertension, and anemia. D: Any treatment for atrial fibrillation or pulmonary hypertension? P: Amiodarone for rhythm control; unsure about pulmonary-hypertension medication.	Collected several essential elements, including ESRD, TAVR, dialysis, medication use, allergy status, hospitalization history, atrial fibrillation, pulmonary hypertension, and anemia. However, atrial fibrillation, pulmonary hypertension, and anemia were elicited late , after broader safety-history questions. Under a fixed-turn cold-start budget, this ordering delays clarification of high-value cardiopulmonary comorbidities.
Physician Annotated Order	D: Age, gender, and chronic diagnoses? P: 73-year-old male; prior valve treatment and kidney failure on dialysis. D: How often is dialysis? P: Three times per week. D: Any other heart, lung, or blood conditions? P: Atrial fibrillation, pulmonary hypertension, and anemia. D: Any related long-term medication? P: Amiodarone; unsure about pulmonary-hypertension medication. D: Any drug allergy, surgery, or hospitalization history? P: No clear drug allergy; prior valve procedure and GI-bleeding hospitalizations.	Best ordering among the five examples. The physician annotation moves atrial fibrillation, pulmonary hypertension, anemia, and related medication use immediately after core comorbidities and dialysis. This better matches the cold-start information-acquisition objective by prioritizing high-value cardiopulmonary and renal evidence before lower-yield background checks.
DeepSeek	D: Age, gender, and chronic diagnoses? P: 73-year-old male; treated severe aortic stenosis and hypertension. D: Any other heart disease? P: Paroxysmal atrial fibrillation. D: Any kidney or lung disease? P: ESRD on regular dialysis; also pulmonary hypertension. D: Any blood thinner for atrial fibrillation? P: No blood thinner. D: What valve procedure was done? P: TAVR with a tissue valve. D: Why no blood thinner? P: Prior serious upper GI bleeding. D: Any blood-pressure or rate-control medication? P: Amlodipine.	Good diagnostic depth for atrial fibrillation, TAVR type, anticoagulation status, pulmonary hypertension, and prior upper GI bleeding. However, it spends turns on anticoagulation reasoning before clarifying broader background such as allergy, hospitalization, or full medication history, suggesting turn-level imbalance .

	Condensed Dialogue	Consultation Evaluation
GPT-5.4	D: Age, gender, and chronic diagnoses? P: 73-year-old male; severe aortic stenosis and hypertension. D: Was the aortic stenosis treated? P: Yes, with TAVR. D: Any remaining valve problem? P: Leakage around the new valve. D: Any other long-term heart condition? P: Atrial fibrillation. D: Any blood thinner or bleeding history? P: No blood thinner; prior upper GI bleeding. D: What heart medication is used? P: Amiodarone for rhythm control.	Strongly follows the aortic-stenosis/TAVR pathway and identifies residual paravalvular leak, atrial fibrillation, bleeding history, and amiodarone. However, it under-elicits ESRD and dialysis early, even though renal replacement therapy is central to longitudinal risk assessment and medication interpretation. The dialogue is coherent but narrower than the physician-reordered trajectory.
MEDCIA	D: Age, gender, and chronic diagnoses? P: 73-year-old male; prior TAVR and ESRD on dialysis. D: How often is dialysis? P: Three times per week. D: Any other heart or blood conditions? P: Atrial fibrillation, heart failure, pulmonary hypertension, and chronic anemia. D: Any remaining valve issue after TAVR? P: Paravalvular aortic leakage. D: What heart, kidney, or GI medications are used? P: Amiodarone, amlodipine, torsemide, sevelamer, and pantoprazole. D: Any procedure, access, allergy, or hospitalization history? P: Prior valve procedure, dialysis access, and GI-bleeding history; no clear drug allergy.	Strong model-generated ordering. The dialogue rapidly captures the highest-value longitudinal risk factors, including ESRD on dialysis, prior TAVR, atrial fibrillation, heart failure, chronic anemia, residual valve leakage, and core cardiovascular/renal medications. Compared with less targeted sequences, it better balances early high-risk evidence acquisition with broader background recovery, while remaining coherent with the fixed-first-question cold-start protocol.