EVOCARE: TOWARDS EFFICIENT AND ROBUST CLINICAL DIAGNOSIS WITH PROACTIVE AND INTERACTIVE SELF-EVOLVING AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical diagnosis starts with incomplete evidence, so the quality of each question shapes the final decision. Yet successful consultations alone reveal little about which questioning strategies will help future patients. We introduce EvoCare, a self-evolving agent that turns consultation histories into reusable inquiry policies. It contrasts related successes and failures to propose targeted rules, then tests each update on a separate cohort before adoption. Across three clinical diagnosis benchmarks and two language-model backbones, EvoCare outperforms interactive and experience-reuse baselines; on StreamBench-DDxPlus, it improves top-rank pathology accuracy by 2.9–6.3 percentage points over the strongest baseline. Further analyses confirm the contribution of each component and show efficient questioning, low repetition, and consistent performance across models. EvoCare points toward a new generation of clinical AI that learns from experience to navigate diagnostic uncertainty more effectively.

1 INTRODUCTION

Clinical diagnosis involves sequential evidence acquisition under incomplete information, with each consultation starting from a partial medical record (Tu et al., 2025; Zhan et al., 2026). Given an underspecified complaint, clinicians elicit the patient's agenda and latent symptoms, formulate and revise differential hypotheses, and determine when the available evidence supports concluding the encounter (Singh Ospina et al., 2019; Tu et al., 2025; Li et al., 2024b). This challenge is increasingly consequential as physician shortages and rising consultation demands strain healthcare systems (Dall et al., 2024), while short primary-care visits (Irving et al., 2017) and limited physical examination in telemedicine (Shaver, 2022) place greater weight on effective history-taking. Proactive and interactive clinical diagnosis remains technically challenging: over multiple turns, a system needs to identify informative questions, interpret responses, update its differential, and terminate efficiently (Li et al., 2024b; Johri et al., 2025; Zhan et al., 2026). In extended clinical dialogues, existing language-model-based systems can lose focus or degrade, especially without explicit inquiry structure (Laban et al., 2026; Hager et al., 2024); these failures matter because shortcomings in history-taking are associated with harmful diagnostic errors (Dalal et al., 2025). This motivates systematic methods for making clinical diagnosis more proactive, interactive, efficient, and robust.

Existing work on clinical diagnosis under incomplete information can be organized into two broad directions: improving diagnosis within an encounter and improving inquiry strategies across encounters. The first direction spans static diagnostic prediction, which operates on findings presented upfront (Jin et al., 2021; Hendrycks et al., 2020) (Figure 1a), and single-encounter interactive dialogue, in which doctor agents acquire evidence through patient interaction (Li et al., 2024b; Schmidgall et al., 2026; Zhan et al., 2026) (Figure 1b). As shown in Figure 1d, diagnosing solely from initial chief complaints significantly lowers scores by 23.0 to 51.2 percentage points relative to complete medical records, which motivates proactive evidence acquisition. The second direction reuses experience across cases through episodic case retrieval (Almansoori et al., 2025) or iterative prompt reflection (He et al., 2026; Agrawal et al., 2026; Yang et al., 2026) (Figure 1c), offering a mechanism for inquiry strategies to improve over time. This field-level view also highlights two unresolved issues in experience reuse: identifying which questioning decisions caused a diagnostic improvement and assessing whether a learned tactic generalizes to a new patient. Our exploratory pre-experiments on a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

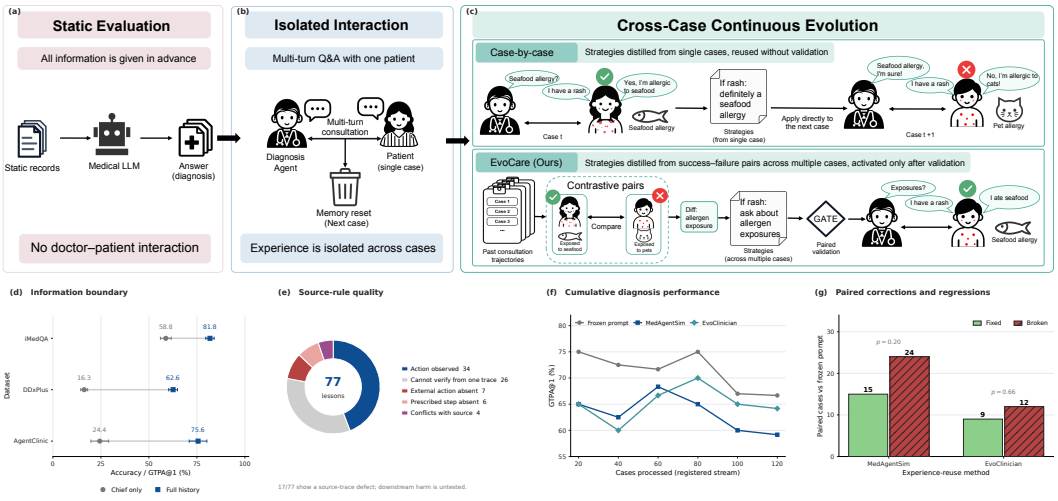


Figure 1: **Diagnostic settings and motivation for cross-case inquiry evolution.** Panels (a–c) show static diagnosis, single-encounter inquiry, and cross-case strategy updates. Panel (d) compares initial-only and full-record diagnosis across three benchmarks; panels (e–g) summarize source-rule attribution and the effects of experience reuse on a 120-case DDxPlus development stream.

registered 120-case clinical stream reveal both issues. First, trajectory-level reflection provides limited attribution: a correct final diagnosis can obscure which specific intermediate question separated competing differentials, motivating heuristics grounded in questions actually observed in the source trace (Figure 1e). Second, clinical heterogeneity can reduce cross-case transfer reliability: in our exploratory stream, two experience-reuse settings produced more regressions than recoveries among previously correct cases (Figure 1f–g).

To address two limitations of experience reuse, namely sparse clinical causality and heterogeneous case-level overfitting, we present EvoCare, a closed-loop policy optimization framework for proactive and interactive clinical diagnosis with self-evolving AI agents. EvoCare separates the roles of patient-facing inquiry and cross-case strategy refinement, decoupling patient interaction from evolutionary learning by externalizing diagnostic strategies into a modular, conditional inquiry policy memory. To disentangle sparse clinical causality, EvoCare introduces cluster-aligned contrastive attribution. By grouping historical consultations into structural behavioral clusters and aligning paired diagnostic successes and failures presenting with homologous symptoms, it identifies intermediate decisions associated with different diagnostic outcomes, proposing conditional inquiry rules grounded in observed trajectory contrasts. To limit case-level overfitting and policy drift, EvoCare uses a cohort-validated clinical safety gate. Following the medical principle of primum non nocere (first, do no harm), each candidate inquiry heuristic is compared with the incumbent policy on an independent cohort of matched patient cases; it is added to the operational policy only if it achieves a strictly higher mean primary score on the separate selection cohort.

We evaluate EvoCare on three multi-turn clinical benchmarks and multiple foundation-model backbones; the results support selection-set-validated inquiry refinement and improved diagnostic yield. Our main contributions are:

- **Problem formulation.** We formalize active clinical inquiry under continuous cross-case evolution, uncovering two previously overlooked phenomena that plague experience reuse: sparse clinical causality and heterogeneous strategy drift that trigger lower diagnostic yield.
- **Methodological design.** We propose EvoCare, a dual-timescale framework that decouples real-time patient inquiry from inter-case heuristic refinement. Cluster-aligned contrastive attribution mines comparable success–failure pairs to identify which questioning decisions actually separate competing differentials, while a cohort-validated clinical safety gate admits each candidate rule only after paired selection-cohort improvement.
- **Evaluation and analysis.** On iMedQA, AgentClinic-MedQA, and StreamBench-DDxPlus, across multiple LLM backbones, and on every reported metric (Table 1), EvoCare consistently establishes

state-of-the-art results against interactive inquiry, episodic retrieval, and prompt/skill optimization baselines. Component ablations confirm that both the attribution and the safety-gate mechanisms are necessary, and the efficiency and robustness comparisons show that EvoCare reaches these results with fewer patient questions, lower learning overhead, and more stable cross-backbone behavior than competing methods.

2 Related Work

Interactive clinical inquiry. Medical QA often presents findings in a single prompt (Jin et al., 2021; Hager et al., 2024). Interactive inquiry lets a doctor agent ask a simulated patient before answering (Li et al., 2024b; Schmidgall et al., 2026; Rose et al., 2025; Du et al., 2025). Prior work examines question quality and within-encounter consistency (Wang et al., 2025; Gong et al., 2025; Laban et al., 2026; Chouayfati et al., 2026; Gong et al., 2026; Werthaim et al., 2026). We study how experience from one encounter changes inquiry policy for later cases.

Cross-case experience and policy evolution. Cross-case approaches reuse earlier encounters through episodic retrieval (Almansoori et al., 2025; Li et al., 2024a; He et al., 2026) or convert trajectories into reusable instructions and skills (Agrawal et al., 2026; Shinn et al., 2023; Zhao et al., 2024; Yang et al., 2026; Zheng et al., 2025; Ni et al., 2026; Sun et al., 2026). EvoCare differs in the unit of evidence it selects: a contrast in observed questioning behavior between comparable cases. It proposes a conditional rule from this contrast and tests the revised policy against the incumbent on separate selection cases before activation.

3 Method

EvoCare separates within-case interaction from cross-case policy revision (Figure 2). The doctor consults each patient with the active skill; offline modules review prior traces, propose a bounded skill edit, and compare candidate and incumbent on separate selection cases.

3.1 Problem Formulation

We formalize active clinical inquiry as an evidence-seeking sequential decision process under a conversational budget B. For a patient i, the agent observes an initial chief complaint $o_{i,0}$. The patient simulator maintains latent clinical findings $\mathcal{E}_i^\star$, while the ground-truth diagnosis y_i is reserved strictly for outcome verification. At dialogue turn $t \le B$, the observable interaction history is:

$$h_{i,t} = \big(o_{i,0}, (q_{i,1}, r_{i,1}), \ldots, (q_{i,t}, r_{i,t})\big), \tag{1}$$

where $q_{i,t}$ denotes the clinician agent's elicited question and $r_{i,t}$ represents the simulator's response conditioned on $\mathcal{E}_i^\star$ and $h_{i,t-1}$.

At each turn, an operational policy $\mathcal{S}$ guides the frozen language model π_θ to emit an action:

$$a_{i,t} \sim \pi_\theta(\cdot \mid h_{i,t}, \mathcal{S}), \quad a_{i,t} \in \{\mathrm{ASK}(q),\ \mathrm{ANSWER}(\hat{y})\}. \tag{2}$$

Selecting $\mathrm{ASK}(q)$ solicits further clinical evidence, while $\mathrm{ANSWER}(\hat{y})$ terminates the consultation with a differential diagnosis. The outcome is assessed by a task-specific diagnostic yield metric $m(\hat{y}_i, y_i) \in [0, 1]$.

The system operates across two decoupled timescales. During intra-encounter inquiry (the fast timescale), the agent asks questions using a fixed policy version $\mathcal{S}$ and the current patient dialogue; weight parameters stay frozen while traces are saved for later learning. Across multiple consultations (the slow timescale), execution traces are aggregated into a trace memory $\mathcal{T}$. Offline optimization modules periodically extract structured heuristics into a pattern repository $\mathcal{P}$, which in turn guides updates to the operational policy $\mathcal{S}$. The operational inquiry policy memory $\mathcal{S} = \{R_1, \ldots, R_K\}$ consists of modular conditional rules, where each rule R specifies an applicability scope, trigger symptom condition, targeted inquiry action, differential update instruction, and exclusion criteria.

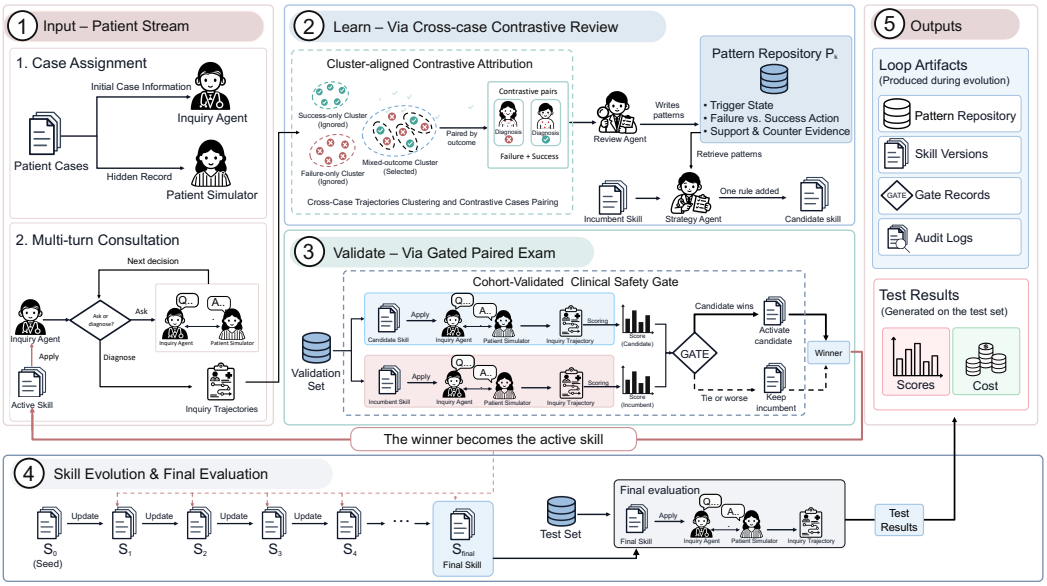


Figure 2: **EvoCare workflow.** The agent consults each patient with an active skill, while contrasts between related consultations inform a candidate rule. Paired selection-set evaluation determines whether the revised skill becomes active for later cases.

3.2 Cluster-aligned Contrastive Attribution

EvoCare compares similar trajectories to identify candidate questions associated with different diagnostic outcomes. These contrasts provide trace-level evidence for selecting a question.

Behavioral distance. For any two completed traces $\tau_i, \tau_j \in \mathcal{T}$, we measure behavioral discrepancy over both dialogue semantics and structural decision paths. Let $\mathbf{v}$ denote the normalized TF-IDF vector of conversational utterances, and let $\mathbf{z}$ represent the chronological sequence of observable clinical actions (e.g., symptom probing types, targeted organ systems, and working differential revisions). The behavioral distance is defined as:

$$\mathcal{D}(\tau_i, \tau_j) = \frac{1}{2}\big[1 - \cos(\mathbf{v}_i, \mathbf{v}_j)\big] + \frac{1}{2}\frac{\mathrm{EditDistance}(\mathbf{z}_i, \mathbf{z}_j)}{\max(|\mathbf{z}_i|, |\mathbf{z}_j|)}. \tag{3}$$

Ground-truth diagnoses and case identifiers are excluded from $\mathbf{v}$ and $\mathbf{z}$. Complete-linkage hierarchical clustering (Johnson, 1967) over $\mathcal{D}$ with a distance threshold partitions historical traces into structural behavioral clusters $\{\mathcal{C}_k\}$.

Comparable trajectory pairs. From each cluster $\mathcal{C}_k$ containing both correct and incorrect outcomes, EvoCare selects a success–failure pair (τ^+, τ^-) with homologous symptoms, operationalized as shared salient terms in their initial complaints. Their related presentations and behavior make observed question differences suitable for review as candidate actions.

Review and skill proposal. A review agent examines each selected pair (τ^+, τ^-), recording what was asked, what the patient reported, and how the working differential changed. It stores candidate evidence gaps and supporting trace IDs in repository $\mathcal{P}$. A strategy agent uses these patterns and the incumbent policy $\mathcal{S}_{k-1}$ to propose one bounded rule edit ΔR, producing a candidate policy:

$$\mathcal{S}'_k = \mathrm{Propose}(\mathcal{S}_{k-1}, \mathcal{P}, \tau^+, \tau^-) = \mathcal{S}_{k-1} \oplus \Delta R. \tag{4}$$

3.3 Cohort-validated Clinical Safety Gate

Each applicable candidate is compared with the incumbent on a selection cohort disjoint from acquisition and test cases. The gate admits a candidate only if its prespecified aggregate score is higher on that cohort; individual selection cases may still regress.

Paired comparison. Let $\mathcal{V}_D$ denote a dedicated validation cohort held disjoint from both acquisition and test sets. In evolutionary round k, candidate policy $\mathcal{S}'_k$ and incumbent policy $\mathcal{S}_{k-1}$ undergo a paired trial on all cases in $\mathcal{V}_D$. Both policies interact with the simulated patients under identical initial complaints, hidden profiles, random seeds, and question budgets B.

Activation criterion. A candidate policy is admitted if and only if it actively triggers on the validation cohort and achieves a strictly superior mean diagnostic score over the incumbent:

$$\mathcal{S}_k = \begin{cases} \mathcal{S}'_k, & \text{if } n_{\text{app}}(\mathcal{S}'_k; \mathcal{V}_D) \geq 1 \quad \text{and} \quad M(\mathcal{S}'_k; \mathcal{V}_D) > M(\mathcal{S}_{k-1}; \mathcal{V}_D), \\ \mathcal{S}_{k-1}, & \text{otherwise}, \end{cases} \tag{5}$$

where $n_{\text{app}}(\mathcal{S}'_k; \mathcal{V}_D)$ counts the number of validation cases where rule ΔR was applied, and $M(\mathcal{S}; \mathcal{V}_D) = \frac{1}{|\mathcal{V}_D|} \sum_{i \in \mathcal{V}_D} m(\hat{y}_i(\mathcal{S}), y_i)$ is the cohort diagnostic yield.

Proposals that fail to trigger, result in performance ties, or degrade cohort accuracy are rejected, retaining $\mathcal{S}_{k-1}$ as the operational policy. The policy active at the conclusion of the evolutionary stream, $\mathcal{S}_{\text{final}}$, is frozen and evaluated once on the unseen test benchmark.

4 Experimental Setups

4.1 Datasets, Tasks, and Evaluation

We define active medical inquiry as the capability of an agent to interactively interview a patient presenting with incomplete clinical information, elicit discriminative findings, and iteratively refine a differential diagnosis under a constrained conversational budget. Acquisition, selection, and held-out evaluation across all benchmarks use strictly disjoint case partitions.

We evaluate EvoCare on three interactive medical benchmarks. Specifically, iMedQA (Jin et al., 2021; Li et al., 2024b) presents abbreviated clinical vignettes and candidate options under an interactive setting, where the agent queries a patient simulator before making a choice, evaluated by diagnostic option accuracy. StreamBench-DDxPlus (Wu et al., 2024; Fansi Tchango et al., 2022) evaluates multi-turn sequential reasoning initialized with patient demographics and chief complaints, requiring the doctor to elicit history and rank five differential candidates; performance is measured by ground-truth pathology accuracy at ranks one and three (GTPA@1 and GTPA@3) as well as the average rank of the ground-truth pathology. AgentClinic-MedQA (Schmidgall et al., 2026) tests free-text consultations with a simulated patient providing history, physical findings, and laboratory results; GTPA@1 and GTPA@3 use a standardized semantic-equivalence judge for clinical synonyms.

4.2 Baseline Methods

We compare EvoCare against four representative baselines: (1) interactive inquiry: MediQ (Li et al., 2024b), which optimizes multi-turn information gathering through adaptive candidate-question selection; (2) episodic retrieval: MedAgentSim (Almansoori et al., 2025), which employs retrieval-based memory over clinical encounters; (3) prompt optimization: GEPA (Agrawal et al., 2026), which refines conversational prompts through reflective textual feedback; and (4) skill revision: SkillOpt (Yang et al., 2026), which applies targeted edits to an operational skill document. All learning-based baselines initialize from the identical clinical context and initial skill document S_0.

4.3 Implementation Details

EvoCare is implemented in Python 3.11 with dependencies managed via uv. All experiments are conducted on a Mac Studio M3 Ultra with 512GB of RAM, with language model backbones accessed via official API endpoints. Experiments employ DeepSeek-V4.1-Flash (Xu et al., 2026) and GLM-5.3-flash (Zeng et al., 2026) as doctor backbones. Models are queried with temperature set to zero and a maximum generation limit of 8,192 tokens. The default interaction budget allows up to five patient inquiries ($B = 5$), with an additional single-turn budget ($B = 1$) evaluated for comparison.

EvoCare processes training cases in batches of ten across up to 28 evolutionary rounds. Complete-linkage clustering combines transcript TF–IDF cosine distance with normalized edit distance over

diagnostic decision sequences at a 0.70 cutoff. Pairwise review selects up to three contrasting success–failure pairs, one per eligible cluster, to propose atomic rule modifications. Candidate skills are adopted only when yielding strict performance improvements on the selection split. We use bootstrapping on all test set samples 10000 times to report the mean and standard deviations.

5 Experimental Results and Analysis

5.1 Main Results: EvoCare Performs Consistently Better Than Baselines

Table 1: **Diagnostic performance across three benchmarks.** We compare EvoCare with MediQ, MedAgentSim, GEPA, and SkillOpt under a five-question budget, reporting DeepSeek-V4.1-Flash and GLM-5.3-flash separately. Scores are multiplied by 100 except for gold-disease rank; the best result in each panel and column is bold.

	iMedQA	**AgentClinic-MedQA**		**StreamBench-DDxPlus**		
Method	Acc. ↑	GTPA@1 ↑	GTPA@3 ↑	GTPA@1 ↑	GTPA@3 ↑	Gold rank ↓
DeepSeek-V4.1-Flash						
MediQ	59.1±2.8	40.7±5.3	60.5±5.3	26.5±2.1	49.7±2.4	3.64±0.10
MedAgentSim	68.9±2.6	55.8±5.4	69.8±5.0	54.2±2.4	69.2±2.2	2.53±0.10
GEPA	66.0±2.6	47.7±5.4	66.3±5.1	52.6±2.4	74.6±2.1	2.37±0.09
SkillOpt	67.9±2.6	57.0±5.3	69.8±4.9	52.2±2.4	67.1±2.3	2.40±0.10
EvoCare	**71.7**±2.5	**61.6**±5.2	**70.9**±4.9	**60.5**±2.3	**74.8**±2.1	**2.19**±0.09
GLM-5.3-flash						
MediQ	71.7±2.5	33.7±5.1	45.3±5.3	55.6±2.4	78.2±2.0	2.31±0.09
MedAgentSim	83.3±2.1	66.3±5.0	76.7±4.9	66.0±2.2	80.3±1.9	1.80±0.08
GEPA	85.2±2.0	64.0±5.2	77.9±4.5	65.1±2.3	82.3±1.8	1.79±0.07
SkillOpt	82.7±2.1	60.5±5.3	73.3±4.8	61.5±2.3	81.9±1.8	1.85±0.07
EvoCare	**86.2**±1.9	**68.6**±5.0	**79.1**±4.4	**68.9**±2.2	**85.7**±1.6	**1.71**±0.07

Efficacy across diverse diagnostic scenarios. As shown in Table 1, EvoCare consistently outperforms interactive inquiry, episodic retrieval, and prompt/skill optimization baselines across all three benchmarks and both foundation backbones. Unlike episodic retrieval (MedAgentSim) or textual prompt adaptation (GEPA), which often yield diffuse clinical guidance, EvoCare dynamically refines modular diagnostic skills. When equipped with DeepSeek-V4.1-Flash, EvoCare achieves leading scores across all evaluation metrics, advancing diagnostic accuracy on iMedQA to 71.7% and exact pathology identification (GTPA@1) on DDxPlus to 60.5%. This superiority is sustained under the GLM-5.3-flash backbone, where EvoCare establishes the highest margins across all tasks. These results confirm that structured policy evolution provides a more reliable mechanism for medical reasoning than unconstrained verbal feedback or instance-level memory lookup.

Discriminative inquiry prioritizes the primary diagnosis. A granular examination of the differential diagnosis metrics on StreamBench-DDxPlus reveals the mechanism driving EvoCare's superiority. While competitive baselines such as GEPA demonstrate comparable capability in capturing the true pathology within a broad candidate pool (74.8% vs. 74.6% GTPA@3 with DeepSeek), EvoCare substantially pulls ahead when isolating the primary diagnosis, posting a 7.9% absolute gain in GTPA@1 (60.5% vs. 52.6%) and promoting the true pathology significantly earlier in the differential list (mean gold rank of 2.19 vs. 2.37). In clinical triage, forming a vague differential list is trivial; the critical challenge is asking targeted, exclusionary questions to eliminate confounders. EvoCare's evolved policy acquires this discriminative inquiry capability, shifting from unfocused clinical questioning to deliberate hypothesis-testing inquiries that elevate the gold diagnosis to the top position.

Closing the empirical information-boundary gap. We compare five-question inquiry with initial-only and full-record diagnosis using the reference-gap coverage in Table 2. EvoCare covers 56.1%, 72.7%, and 95.5% of the reference gap on iMedQA, AgentClinic-MedQA, and StreamBench-DDxPlus, respectively, exceeding the strongest baseline in each task by 12.2, 9.0, and 13.6 percentage

Table 2: **Five-question diagnostic reference-gap coverage.** With DeepSeek-V4.1-Flash, Initial only and Full record diagnose without patient questions. Coverage is $R_D = 100(M_{B=5} - M_{\text{initial}})/(M_{\text{full}} - M_{\text{initial}})$; reference scores are mean ± SD, and coverage uses the reported score estimates. The best coverage per benchmark is bold.

	iMedQA	**AgentClinic-MedQA**	**StreamBench-DDxPlus**
Setting / Method	Acc.	GTPA@1	GTPA@1
Diagnostic score (%)			
Initial only	58.8±2.8	24.4±4.6	16.3±1.8
Full record	81.8±2.2	75.6±4.6	62.6±2.3
Reference-gap coverage R_D (%) ↑			
MediQ	1.3	31.8	22.0
MedAgentSim	43.9	61.3	81.9
GEPA	31.3	45.5	78.4
SkillOpt	39.6	63.7	77.5
EvoCare	**56.1**	**72.7**	**95.5**

points. On DDxPlus, its five-question GTPA@1 of 60.5% approaches the full-record reference score of 62.6% from an initial-only score of 16.3%.

5.2 Ablation Study: Cluster-aligned Contrastive Attribution and the Cohort-validated Clinical Safety Gate Enhance Accuracy and Safety

StreamBench-DDxPlus yields the most pronounced information gain and multi-turn clinical complexity, so we use it to analyze EvoCare's policy updates. We ablate two mechanisms: *No aligned selection* removes cluster-aligned contrastive attribution by omitting behavioral clustering and chief-complaint alignment, while *Direct activation* bypasses the paired selection-cohort validation of the cohort-validated clinical safety gate (Table 3).

Table 3: **Component ablations on StreamBench-DDxPlus.** Under DeepSeek-V4.1-Flash, *No aligned selection* removes cluster-aligned contrastive attribution (behavioral clustering and chief-complaint alignment); *Direct activation* bypasses the cohort-validated clinical safety gate (paired selection-cohort validation). Learning tokens count optimizer use, and Repaired/Broken compare each variant with the frozen initial skill. Best diagnostic and case-transition results are bold.

Policy Variant	**GTPA@1** ↑	**Learning Tokens** ↓	**Repaired** ↑	**Broken** ↓
EvoCare	**60.5**±2.3	0.94 M	**72**	**28**
No aligned selection	55.6±2.4	1.74 M	65	43
Direct activation	37.9±2.3	0.46 M	42	98

Cluster-aligned selection improves accuracy and learning efficiency. Removing behavioral clustering and complaint alignment (*No aligned selection*) lowers GTPA@1 from 60.5% to 55.6% and increases optimizer token use from 0.94M to 1.74M (+85%). The aligned success–failure pairs focus rule proposals on observed questioning differences between related presentations, improving rank-one diagnosis with lower optimizer token use.

The cohort-validated gate preserves more correct diagnoses. Bypassing the cohort-validated clinical safety gate (*Direct activation*) lowers held-out GTPA@1 from 60.5% to 37.9%, a 22.6-percentage-point drop. Relative to the frozen initial skill, Direct activation repairs 42 cases and breaks 98 (net -56), whereas EvoCare repairs 72 and breaks 28 (net $+44$). The gate compares candidate and incumbent on the same separate selection cohort and admits a policy edit only after a strictly higher mean primary score.

5.3 Efficiency Analysis: EvoCare Uses Limited Questions with Low Repetition

We next evaluate how efficiently EvoCare interacts under strict budget constraints, examining diagnostic elasticity, conversational discipline, and data-efficiency during policy acquisition.

Diagnostic elasticity across question budgets. Figure 3 illustrates model performance under stringent budget restrictions. Even when constrained to a single question ($B = 1$), EvoCare retains top diagnostic accuracy on iMedQA (66.4%) and AgentClinic-MedQA (43.0%), demonstrating that its evolved skills prioritize high-entropy pivot questions immediately. As the inquiry allowance expands to five questions ($B = 5$), EvoCare demonstrates the strongest diagnostic elasticity, surging by 29.9% on DDxPlus to reach 60.5% GTPA@1. This steep scaling curve confirms that EvoCare does not merely ask arbitrary questions to fill conversational slots; rather, each subsequent round builds upon prior evidence to progressively resolve diagnostic ambiguity.

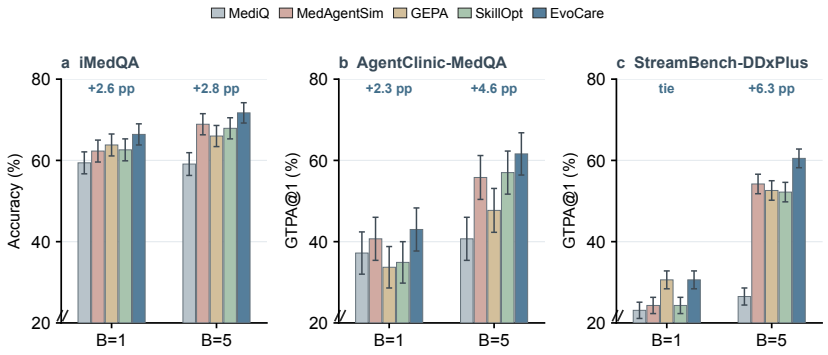


Figure 3: **Diagnostic performance under one- and five-question budgets.** On iMedQA, AgentClinic-MedQA, and StreamBench-DDxPlus, all five methods use DeepSeek-V4.1-Flash with at most one or five patient questions.

Table 4: **Inquiry behavior on AgentClinic-MedQA and StreamBench-DDxPlus.** All methods use DeepSeek-V4.1-Flash and a five-question budget. Budget-cap and zero-question rates are per case; exact repeats are per question asked. The lowest zero-question and exact-repeat rates within each benchmark are bold.

Benchmark	Method	Budget cap (%)	Zero questions (% ↓)	Exact repeats (% ↓)
AgentClinic-MedQA	MediQ	81.4	1.16	11.78
	MedAgentSim	95.3	1.16	0.71
	GEPA	100.0	**0.00**	0.93
	SkillOpt	96.5	**0.00**	0.71
	EvoCare	73.3	**0.00**	**0.51**
StreamBench-DDxPlus	MediQ	8.6	26.76	5.14
	MedAgentSim	98.9	**0.00**	1.73
	GEPA	99.8	**0.00**	0.32
	SkillOpt	99.1	0.91	0.41
	EvoCare	88.0	**0.00**	**0.00**

Inquiry discipline and autonomous stopping. Clinical conversational metrics in Table 4 reveal distinct behavioral traits. Baseline agents systematically exhaust their inquiry quota regardless of clinical necessity, hitting the five-question cap in 95.3%–100.0% of cases. In sharp contrast, EvoCare concludes consultation early in 26.7% of AgentClinic-MedQA cases and 12.0% of DDxPlus cases without any sacrifice in diagnostic precision. EvoCare asks at least one question in every evaluated AgentClinic-MedQA and DDxPlus case, whereas MediQ has a 26.76% zero-question rate on DDxPlus. EvoCare records an exact-repeat rate of 0.00% on DDxPlus and 0.51% on AgentClinic-

MedQA. These results show active questioning, earlier stopping than most baselines, and low exact repetition at the observed diagnostic accuracy.

Early gains with partial acquisition exposure. We examine the trajectory of policy acquisition across partial exposures of the training stream (Table 5). EvoCare exhibits exceptional learning sample efficiency: with exposure to merely 25% of the acquisition stream, it reaches 72.6% on iMedQA and 56.9% on DDxPlus, already surpassing the performance of GEPA and SkillOpt trained on 100% of the data (67.9% and 52.6%, respectively). At 50% exposure, AgentClinic-MedQA reaches its full-stream score of 61.6%, while DDxPlus rises further from 59.4% at 50% to 60.5% at 75%. These results show that EvoCare learns useful inquiry rules from partial acquisition streams, with gains appearing before full exposure.

Table 5: **Diagnostic performance across acquisition exposure levels.** GEPA, SkillOpt, and EvoCare learn from 25%, 50%, 75%, or 100% of the acquisition cases under DeepSeek-V4.1-Flash. The highest result for each benchmark and exposure level is bold.

Benchmark / Metric	Method	25% Stream	50% Stream	75% Stream	100% Stream
iMedQA / Acc. ↑	GEPA	70.8±2.5	70.4±2.5	67.9±2.6	66.0±2.6
	SkillOpt	66.4±2.6	69.2±2.6	67.0±2.6	67.9±2.6
	EvoCare	**72.6**±2.5	**70.8**±2.5	**71.7**±2.5	**71.7**±2.5
AgentClinic / GTPA@1 ↑	GEPA	50.0±5.4	47.7±5.4	53.5±5.4	47.7±5.4
	SkillOpt	46.5±5.4	44.2±5.4	39.5±5.3	57.0±5.3
	EvoCare	**54.6**±5.4	**61.6**±5.2	**61.6**±5.2	**61.6**±5.2
DDxPlus / GTPA@1 ↑	GEPA	51.7±2.4	49.4±2.4	56.0±2.4	52.6±2.4
	SkillOpt	54.2±2.4	51.0±2.4	56.0±2.4	52.2±2.4
	EvoCare	**56.9**±2.4	**59.4**±2.3	**60.5**±2.3	**60.5**±2.3

5.4 Robustness Evaluation: EvoCare Retains Leading Results Across Backbones

A fundamental requirement for clinical agent frameworks is resilience to varying foundation model backbones. As reported in Table 1, EvoCare's architectural advantages transcend the choice of underlying LLM. When migrated from DeepSeek-V4.1-Flash to GLM-5.3-flash, EvoCare continues to outperform all competing frameworks across every evaluation metric, scaling to 86.2% on iMedQA, 68.6% on AgentClinic-MedQA, and an outstanding 1.71 gold pathology rank on DDxPlus.

More importantly, EvoCare stabilizes multi-turn interactive reasoning across varying model architectures. On AgentClinic-MedQA, baseline frameworks display volatile shifts under backbone migration; for example, MediQ's GTPA@1 collapses from 40.7% to 33.7% when driven by GLM-5.3-flash due to degraded conversational multi-turn coordination. In contrast, EvoCare's modular policy structure provides robust procedural guidance that absorbs backbone-level variances, consistently advancing performance (from 61.6% to 68.6%). These results show that EvoCare retains the highest reported scores across both evaluated LLM backbones.

6 Conclusion

In this paper, we introduced EvoCare, a self-evolving agent framework that improves clinical inquiry across patient encounters. EvoCare turns consultation experience into actionable strategy: contrasts between related successful and unsuccessful trajectories reveal candidate questioning rules, and paired evaluation on a separate cohort determines which policy updates are adopted. This design lets the agent refine its strategy across cases while following a stable policy within each consultation. Across iMedQA, AgentClinic-MedQA, and StreamBench-DDxPlus, EvoCare achieves the best results among the evaluated methods on every reported diagnostic metric with both language-model backbones. Component ablations show that aligned trajectory selection and cohort validation each contribute to these gains. Together, the results support targeted policy updates drawn from observed decisions and retained only when they improve cross-case diagnostic performance.

AI USE STATEMENT

In this work, we used generative AI tools for the implementation of the EvoCare framework. We have not used generative AI tools for proposing or refining hypotheses or for designing the research methodology, and the remaining required tasks—generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, translating, cleaning or reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results—are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, including the Python code that renders the figures and tables from our recorded experimental results, for editing the paper to improve readability, including the LaTeX layout of the manuscript, and for brainstorming the title and keywords. We have reviewed all AI-assisted work: the authors designed the methodology and chose the experimental parameters and thresholds, and we verified the code by running the pipeline and reproducing the reported results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments in this work use the publicly released, de-identified benchmark datasets iMedQA, AgentClinic-MedQA, and StreamBench-DDxPlus, under their respective licenses and intended research use. The work does not involve human participants and collects no new patient data, and no personally identifiable or protected health information is used at any stage.

REPRODUCIBILITY STATEMENT

We provide the complete codebase as supplementary material, and the Appendix documents the dataset partitions, evaluation metrics and uncertainty, question budgets, the policy-evolution procedure, and the clustering-threshold sensitivity analysis.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8479–8565, 2026.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Medagentsim: Self-evolving multi-agent simulations for realistic clinical interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 362–372. Springer, 2025.

Pia Chouayfati, Alexander M Fichtl, Miriam Anschütz, George Doumat, and Georg Groh. Mtdiag: A multi-turn diagnostic dataset towards clinically meaningful llm evaluation. In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 821–840, 2026.

Anuj K Dalal, Savanna Plombon, Kaitlyn Konieczny, Daniel Motta-Calderon, Maria Malik, Alison Garber, Alyssa Lam, Nicholas Piniella, Marie Leeson, Pamela Garabedian, et al. Adverse diagnostic events in hospitalised patients: a single-centre, retrospective cohort study. *BMJ Quality & Safety*, 34(6):377–388, 2025.

Tim Dall, Ryan Reynolds, Ritashree Chakrabarti, Clark Ruttinger, Patrick Zarek, and Owen Parker. The complexities of physician supply and demand: projections from 2021 to 2036. *Association of American Medical Colleges*, 2024.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. Llms can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17278–17306, 2025.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35: 31306–31318, 2022.

Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, et al. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*, 2026.

Linlu Gong, Ante Wang, Yunghwei Lai, Weizhi Ma, and Yang Liu. The dialogue that heals: A comprehensive evaluation of doctor agents' inquiry capability. *arXiv preprint arXiv:2509.24958*, 2025.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Yufei He, Juncheng Liu, Zhiyuan Hu, Yulin Chen, Yue Liu, Yuan Sui, Yibo Li, Nuo Chen, Jun Hu, Bryan Hooi, et al. Evoclinician: A self-evolving agent for multi-turn medical diagnosis via test-time evolutionary learning. *arXiv preprint arXiv:2601.22964*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Greg Irving, Ana Luisa Neves, Hajira Dambha-Miller, Ai Oishi, Hiroko Tagashira, Anistasiya Verho, and John Holden. International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ open*, 7(10):e017902, 2017.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://doi.org/10.3390/app11146421.

Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86, 2025.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024a.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024b.

Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158*, 2026.

Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13803–13826, 2025.

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

Julia Shaver. The state of telehealth before and after the covid-19 pandemic. *Primary care*, 49(4): 517, 2022.

N Shinn, F Cassano, E Berman, A Gopinath, K Narasimhan, and SR Yao. Language agents with verbal reinforcement learning. *arXiv*, 2023.

Naykky Singh Ospina, Kari A Phillips, Rene Rodriguez-Gutierrez, Ana Castaneda-Guarderas, Michael R Gionfriddo, Megan E Branda, and Victor M Montori. Eliciting the patient's agenda-secondary analysis of recorded clinical encounters. *Journal of general internal medicine*, 34(1): 36–40, 2019.

Haoran Sun, Wenjie Li, Yujie Zhang, Zekai Lin, Fanrui Zhang, Kaitao Chen, Xingqi He, Yichen Li, Mianxin Liu, Lei Liu, et al. Experience makes skillful: Enabling generalizable medical agent reasoning via self-evolving skill memory. *arXiv preprint arXiv:2606.09365*, 2026.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025. doi: 10.1038/s41586-025-08866-7.

Ziyu Wang, Hao Li, Di Huang, Hye-Sung Kim, Chae-Won Shin, and Amir M Rahmani. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations. *Smart Health*, 36: 100570, 2025.

Mai Werthaim, Maya Kimhi, Alexander Apartsin, and Yehudit Aperstein. A benchmark for evaluating diagnostic questioning efficiency of llms in patient conversations. *Scientific Reports*, 16(1):6121, 2026.

Cheng-Kuang Wu, Zhi R Tam, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. Streambench: Towards benchmarking continuous improvement of language agents. *Advances in Neural Information Processing Systems*, 37:107039–107063, 2024.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, et al. Skillopt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Chen Zhan, Xihe Qiu, Xiaoyu Tan, Xibing Zhuang, Gengchen Ma, Yue Zhang, Shuo Li, Peifeng Liu, Xiaoxiao Ge, Liang Liu, et al. Active evidence-seeking and diagnostic reasoning in large language models for clinical decision support. *arXiv preprint arXiv:2605.22047*, 2026.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025.

A Datasets, Splits, and Protocol Attribution

Evaluations use standardized, frozen dataset splits to ensure fair comparison across all baselines.

Table 6: **Dataset partitions for the three benchmarks.** Acquisition cases support policy updates, selection cases determine candidate admission, and test cases provide final scores.

Benchmark	Acquisition	Selection	Test	Partition
iMedQA	254	64	318	25% group-stratified sample
StreamBench-DDxPlus	274	69	441	25% group-stratified sample
AgentClinic-MedQA	86	42	86	Deterministic 2:1:2 partition

B Question-Budget Results

Table 7: **Exact diagnostic results at one- and five-question budgets.** All five methods use DeepSeek-V4.1-Flash on the same benchmark test cohorts. Reference scores are mean ± SD, multiplied by 100; bold marks the highest point estimate for each benchmark and budget, including the tie on StreamBench-DDxPlus at $B = 1$.

Benchmark / Metric	Method	$B = 1$	$B = 5$
iMedQA / Accuracy ↑	MediQ	59.4±2.7	59.1±2.8
	MedAgentSim	62.3±2.7	68.9±2.6
	GEPA	63.8±2.7	66.0±2.6
	SkillOpt	62.6±2.7	67.9±2.6
	EvoCare	**66.4**±2.6	**71.7**±2.5
AgentClinic-MedQA / GTPA@1 ↑	MediQ	37.2±5.2	40.7±5.3
	MedAgentSim	40.7±5.3	55.8±5.4
	GEPA	33.7±5.1	47.7±5.4
	SkillOpt	34.9±5.1	57.0±5.3
	EvoCare	**43.0**±5.3	**61.6**±5.2
StreamBench-DDxPlus / GTPA@1 ↑	MediQ	23.1±2.0	26.5±2.1
	MedAgentSim	24.3±2.0	54.2±2.4
	GEPA	**30.6**±2.2	52.6±2.4
	SkillOpt	24.3±2.0	52.2±2.4
	EvoCare	**30.6**±2.2	**60.5**±2.3

C Implementation and Sensitivity

Policy-evolution procedure. The active Skill $\mathcal{S}_{k-1}$ remains fixed while each acquisition batch is interviewed. A Review Agent compares selected success–failure traces; a Strategy Agent proposes a bounded edit to the Skill. The candidate is activated only after it applies to at least one selection case and improves the mean primary diagnostic score over the incumbent on the same selection cohort.

Algorithm 1 Cross-case evolution of the inquiry Skill. Behavioral clustering precedes chief-complaint matching within eligible clusters; candidate admission uses a disjoint, paired selection cohort.

Require: Initial Skill $\mathcal{S}_0$; acquisition batches $\{\mathcal{A}_k\}$; disjoint selection cohort $\mathcal{V}_D$; question budget B.

Initialize trace memory $\mathcal{T} \leftarrow \emptyset$ and pattern repository $\mathcal{P} \leftarrow \emptyset$.
for each acquisition batch $\mathcal{A}_k$ **do**
  Set $\mathcal{S}_k \leftarrow \mathcal{S}_{k-1}$; run $\mathcal{S}_{k-1}$ on $\mathcal{A}_k$ under budget B and add traces to $\mathcal{T}$.
  Cluster traces by behavioral distance; within mixed-outcome clusters, select up to three success–failure pairs with shared salient initial-complaint terms.
  if at least one comparable pair is available **then**
    Review observed question and outcome contrasts; update $\mathcal{P}$ and propose one bounded-rule candidate $\mathcal{S}'_k$.
    if the candidate rule applies on $\mathcal{V}_D$ **then**
      Evaluate $\mathcal{S}'_k$ and $\mathcal{S}_{k-1}$ on the same cases in $\mathcal{V}_D$.
      if $M(\mathcal{S}'_k; \mathcal{V}_D) > M(\mathcal{S}_{k-1}; \mathcal{V}_D)$ **then**
        Activate $\mathcal{S}_k \leftarrow \mathcal{S}'_k$.
      end if
    end if
  end if
end for
Return the final frozen Skill $\mathcal{S}_{\text{final}}$ for held-out testing.

Clustering threshold. The complete-linkage sensitivity scan in Figure 4 compares contrastable clusters with success–failure pairs delivered to Review under a three-pair budget. At the adopted threshold $\tau = 0.70$, 49 clusters support outcome contrasts (98% of the observed maximum at $\tau = 0.65$), and 2.3 pairs reach Review per round, 0.1 more than at $\tau = 0.65$.

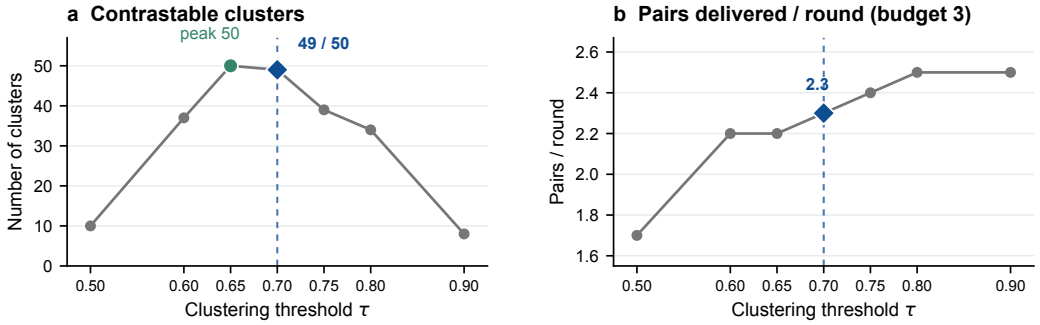


Figure 4: **Sensitivity to the behavioral-clustering distance threshold.** The left panel counts clusters containing both successful and unsuccessful trajectories; the right panel shows aligned success–failure pairs delivered to Review per round under a three-pair budget. The dashed line marks the adopted $\tau = 0.70$ (49 contrastable clusters and 2.3 pairs per round). The plot shows the complete seven-point reported scan.