ClinicalSkill: Explicit Clinical Action Policy for Self-Evolving Medical Agents

Anonymous ACL submission

Abstract

Medical agents are expected to move beyond static question answering toward multi-turn diagnosis and treatment, where decisions must be grounded in evolving clinical evidence and safety constraints. However, existing self-evolving agents remain insufficient for this setting: 1) Retrospective memory: retrieved cases describe past experience but do not provide explicit rules for future actions; 2) Opaque latent policy: experience internalized in model behavior is difficult to inspect, revise, and constrain. To address these issues, we propose ClinicalSkill, an explicit clinical action policy framework for self-evolving medical agents, which transforms completed clinical interactions into governable action policies. Specifically, ClinicalSkill distills diagnosis-treatment trajectories into structured skill cards and maintains them in a reusable skill bank, allowing the agent to make evidence-grounded and safety-aware decisions in future cases. Experiments on multi-disease diagnosis-treatment scenarios show that ClinicalSkill consistently outperforms strong baselines, improving diagnostic and treatment performance by 17.36% and 5.46%, respectively.

1 Introduction

Large language model agents are moving from static response generation toward interactive systems that act, observe, and adapt through experience (Gao et al., 2025; Liu et al., 2025; Chhikara et al., 2025; Yan et al., 2025; Zhang et al., 2026; Yang et al., 2026b). This shift is particularly important in medicine, where clinical decision-making is not a single-shot prediction task but a sequential process under uncertainty. A medical agent must acquire evidence, update the diagnostic state, and recommend treatment under safety constraints. Therefore, medical self-evolution should not only improve final answers, but also produce reusable and governable decision rules for future clinical actions.

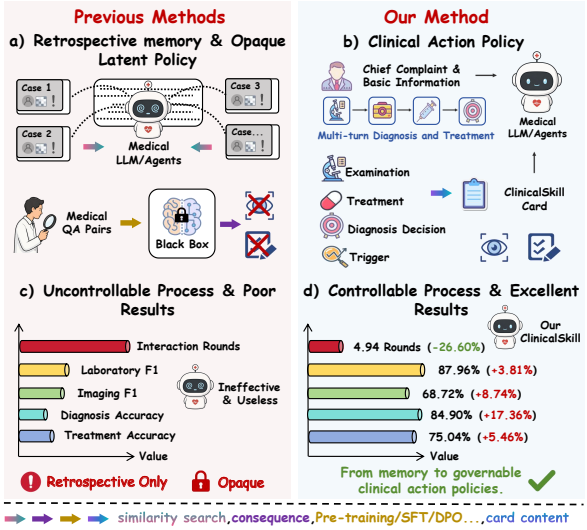

Figure 1: Existing self-evolving agents either store past experience as memory or hide learned behavior in latent policies. ClinicalSkill instead learns explicit clinical action policies to govern future examination, diagnosis, and treatment decisions.

Recent studies have begun to evaluate and improve LLM agents in interactive medical scenarios. Benchmarks such as SDBench, MedGUIDE, MedAgentsBench, and MedMT-Bench move beyond static medical question answering toward interactive clinical decision-making that involves diagnosis, treatment, and long-horizon reasoning (Nori et al., 2025; Tang et al., 2025; Li et al., 2025; Fang et al., 2026; Yang et al., 2026a). Meanwhile, self-evolving medical agents such as EvoClinician, DxEvolve, EvoMDT, and Evo-MedAgent show that interaction histories can improve evidence acquisition and clinical reasoning through test-time adaptation (He et al., 2026; Ren et al., 2026; Liu et al., 2026; Shen et al., 2026). These works demonstrate that clinical agents should learn from interaction histories rather than rely only on static model knowledge. However, most existing systems still treat experi-

ence as diagnostic memory or implicit adaptation, rather than as an explicit policy object that governs future clinical actions.

This limitation becomes critical for high-risk diagnosis-treatment decision-making. As illustrated in Figure 1, existing self-evolving agents face two key limitations. First, memory-based agents are retrospective: they can retrieve what happened in similar past cases, but they do not define how future clinical actions should be selected and constrained under a new patient state. Second, latent policy adaptation is opaque: it may change model behavior through internalized experience, but the learned decision rules remain difficult to inspect or revise. These limitations suggest that medical self-evolution should not merely store past experience or hide adaptation inside model behavior. In high-risk medicine, self-evolution cannot be an uncontrolled accumulation of memories or an opaque drift in model behavior. It must produce a representation that is prospective rather than retrospective, explicit rather than latent, executable rather than descriptive, and maintainable rather than fixed.

To address these issues, we propose ClinicalSkill, a self-evolving framework that distills multi-turn diagnosis-treatment trajectories into explicit clinical action policies. We design a structured skill-card representation to convert episodic clinical experience into actionable policy units, where each skill specifies when it applies, what evidence should be acquired, how diagnosis should be supported, and how treatment should be recommended. After each completed episode, a rule-guided evaluator assesses the trajectory and produces corrected clinical feedback; a controller synthesizes the evaluated trajectory into a structured skill card; and an inspector performs leakage-aware validation and skill fusion to maintain the ClinicalSkill Bank. Through this closed loop, ClinicalSkill transforms episodic clinical experience into auditable, maintainable, and executable skill units that guide future examination, diagnosis, and treatment decisions. Our contributions are summarized as follows:

- **First Exploration.** To the best of our knowledge, we are the first to formulate medical agent self-evolution in multi-turn diagnosis-treatment scenarios as explicit clinical action policy learning, moving beyond retrospective memory and opaque latent adaptation.
- **Novel Framework.** We propose ClinicalSkill, a closed-loop self-evolving framework that converts clinical trajectories into reusable skill cards for governing future clinical actions.
- **SOTA Performance.** Extensive experiments on the MIMIC-CDM diagnosis-treatment benchmark show that ClinicalSkill consistently outperforms vanilla LLM agents and the strong DxEvolve baseline, improving diagnosis, treatment, evidence acquisition, and interaction efficiency.

2 Related Work

2.1 Self-Evolving Agents

Recent self-evolving agents mainly improve through external memory or parameter internalized policy. Memory-based methods store trajectories, reflections, or episodic traces for later retrieval (Liu et al., 2025; Chhikara et al., 2025; Wei et al., 2025). More adaptive systems, such as Memory-R1, and MemSkill, further learn to manage and update memory (Yan et al., 2025; Zhang et al., 2026).These methods enable lightweight adaptation, but memory remains retrospective: it records what happened before, without explicitly defining what should be done next. In contrast, RL- or parameter-based self-improvement can internalize experience into model behavior (Yan et al., 2025; Wang et al., 2025), but the resulting policy is often latent and difficult to inspect or edit. ClinicalSkill differs by transforming interaction trajectories into explicit clinical action policies, specifying when an action applies, why it is justified, and under which constraints it should be blocked or revised.

2.2 Interactive Medical Agents

Medical LLM evaluation has recently moved from static question answering toward sequential and multi-turn clinical decision-making. Benchmarks such as SDBench, MINT, MedGUIDE, MedAgentsBench, and MedMT-Bench evaluate interactive clinical decision-making across diagnosis, treatment, and long-horizon reasoning (Nori et al., 2025; Fang et al., 2026; Li et al., 2025; Tang et al., 2025; Yang et al., 2026a). Recent medical agents further improve clinical reasoning through test-time adaptation, experience reuse, and agent-level feedback (He et al., 2026; Ren et al., 2026; Liu et al., 2026; Shen et al., 2026). These works

show that clinical intelligence cannot be fully measured by static final-answer accuracy, but most of them still represent accumulated experience as diagnostic memory, cognition primitives, or evolving heuristics. In contrast, ClinicalSkill targets the full diagnosis-treatment chain and learns explicit clinical action policies that guide future examination, diagnosis, and treatment decisions under safety constraints.

2.3 Skill Learning and Policy Abstraction

Skill-based agents convert repeated interactions into reusable behavioral units. Recent works such as AutoSkill, MemSkill, XSkill, and skill-library reinforcement learning show that skill abstraction can support reuse, maintenance, and transfer beyond raw trajectory storage (Yang et al., 2026b; Zhang et al., 2026; Jiang et al., 2026; Wang et al., 2025). However, most existing skill formulations are designed for general task-solving and tool-use settings, where a skill can often be treated as a reusable instruction or task template. High-risk clinical decision-making requires stricter policy abstraction: a clinical skill must encode applicability conditions, evidence requirements, diagnostic criteria, treatment constraints, and safety boundaries. In contrast, ClinicalSkill formulates each skill as an auditable and executable clinical action policy, retrieved not to recall a similar case but to constrain what the agent should examine, diagnose, and treat under the current patient state.

3 Methods

We present ClinicalSkill, a self-evolving framework that distills multi-turn diagnosis-treatment interactions into explicit clinical action policies. As shown in Figure 2, the framework consists of four stages: multi-turn diagnosis and treatment, rule-guided trajectory evaluation, skill card synthesis, and leakage-aware skill bank maintenance.

3.1 Multi-turn Diagnosis and Treatment

We formulate diagnosis-treatment interaction as a sequential clinical decision process. At turn t, the agent observes the patient state x_t, maintains the trajectory history h_t, selects a clinical action a_t, and receives a new observation o_t. The action space is decomposed as:

$$\mathcal{A} = \mathcal{A}_{\text{exam}} \cup \mathcal{A}_{\text{diag}} \cup \mathcal{A}_{\text{tx}}, \tag{1}$$

where $\mathcal{A}_{\text{exam}}$ contains examination requests, $\mathcal{A}_{\text{diag}}$ contains diagnostic decisions, and $\mathcal{A}_{\text{tx}}$ contains treatment recommendations. A completed episode is denoted as:

$$\tau = (x_0, a_0, o_0, \ldots, x_T, d, y), \tag{2}$$

where d is the final diagnosis and y is the final treatment recommendation.

ClinicalSkill does not store τ as a retrospective case memory. Instead, it distills clinical experience into reusable action-policy units. We maintain a ClinicalSkill Bank $\mathcal{B} = \{s_i\}_{i=1}^{N}$, where each skill follows the schema:

$$s_i = (t_i, e_i, r_i, u_i, \pi_i, \mu_i). \tag{3}$$

Here, t_i is the trigger module, e_i is the examination policy, r_i is the diagnostic decision rule, u_i is the treatment policy, π_i is the runtime prompt, and μ_i stores provenance and maintenance metadata. This schema makes a ClinicalSkill different from a retrieved case or reflection: each skill explicitly specifies when it should be activated, what evidence should be acquired, how a diagnosis should be reached, and how treatment should be recommended under safety constraints.

Before each decision, the agent builds a trigger query $q_t = g(x_t, h_t)$ from symptoms, temporal cues, abnormal findings, negative evidence, and diagnostic uncertainty. The active skill set is retrieved as:

$$\mathcal{S}_t = \text{Retrieve}_K(q_t, \mathcal{B}), \tag{4}$$

where retrieval ranks skills by matching q_t with their trigger modules using hybrid semantic and lexical matching. The agent then selects the next clinical action by conditioning on the patient state, trajectory history, and retrieved skill prompts:

$$a_t = \mathcal{M}_{\text{agent}}\left(x_t, h_t, \{\pi_s : s \in \mathcal{S}_t\}\right). \tag{5}$$

After executing a_t, the environment returns o_t, the patient state is updated, and the interaction continues until the agent outputs both a final diagnosis and a treatment recommendation, or reaches the maximum turn budget.

3.2 Trajectory Evaluation

After an episode is completed, ClinicalSkill evaluates the interaction trajectory before using it for skill construction. The evaluator consists of an evaluation rulebook $\mathcal{R}_{\text{eval}}$ and an evaluator LLM

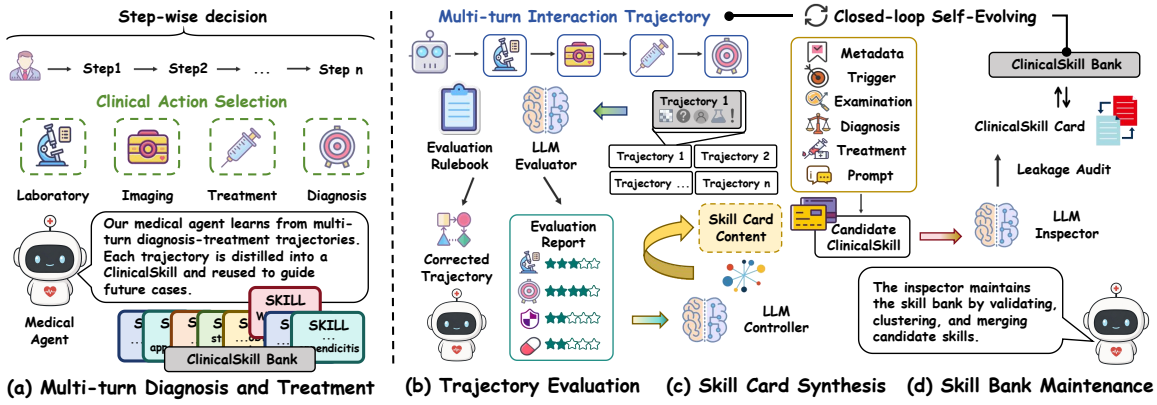

Figure 2: Overview of ClinicalSkill. The medical agent performs multi-turn diagnosis and treatment with retrieved skills, while completed trajectories are evaluated, corrected, and distilled into structured ClinicalSkill cards. An inspector maintains the ClinicalSkill Bank through validation, clustering, and merging, enabling a closed-loop self-evolving process for future clinical cases.

$\mathcal{M}_{\text{eval}}$. Given the completed trajectory τ and reference information z, the evaluator produces an evaluation report ϵ and a corrected trajectory $\tilde{\tau}$:

$$(\epsilon, \tilde{\tau}) = \mathcal{M}_{\text{eval}}\left(\tau, z, \mathcal{R}_{\text{eval}}\right). \tag{6}$$

The evaluator is used only after an episode is completed for skill construction; it does not intervene in the agent's step-by-step decision process during held-out evaluation.

The evaluation report ϵ assesses, under the predefined rules, whether examination actions are necessary and well-ordered, whether the final diagnosis is sufficiently supported, whether the treatment recommendation is consistent with the diagnosis, and whether unsafe or unsupported actions appear in the trajectory. The corrected trajectory $\tilde{\tau}$ marks unsupported actions, annotates missing evidence, revises diagnostic reasoning, and aligns treatment decisions with the evaluated diagnosis and safety constraints. This step prevents the controller from distilling skills directly from noisy or clinically flawed trajectories.

3.3 Skill Card Synthesis

The controller converts the evaluator outputs into a draft ClinicalSkill card. Let $\mathcal{R}_{\text{skill}}$ denote the predefined skill-card schema. The controller LLM $\mathcal{M}_{\text{ctrl}}$ generates:

$$\hat{s} = \mathcal{M}_{\text{ctrl}}\left(\tilde{\tau}, \epsilon, \mathcal{R}_{\text{skill}}\right), \tag{7}$$

$$\hat{s} = (\hat{t}, \hat{e}, \hat{r}, \hat{u}, \hat{\pi}, \hat{\mu}). \tag{8}$$

Here, $\hat{s}$ is a draft ClinicalSkill card. The controller is not used to produce a free-form trajectory summary for direct reuse. Instead, its output is constrained by the ClinicalSkill schema, which requires separate trigger, examination, diagnosis, treatment, prompt, and metadata modules.

Each module captures a different part of the clinical action policy within the card. The trigger module $\hat{t}$ specifies applicability and exclusion conditions. The examination module $\hat{e}$ specifies evidence acquisition actions, escalation rules, and stop criteria. The diagnostic module $\hat{r}$ encodes supporting evidence, refuting evidence, differential diagnoses, and diagnostic commitment criteria. The treatment module $\hat{u}$ specifies primary treatment, supportive care, contraindication checks, and monitoring suggestions. The metadata $\hat{\mu}$ records provenance, evaluated outcome, correctness signal, and maintenance history.

The runtime prompt is compiled from the structured clinical modules:

$$\hat{\pi} = \text{Compile}\left(\hat{t}, \hat{e}, \hat{r}, \hat{u}, \Gamma\right), \tag{9}$$

where Γ denotes global guardrails, including evidence-first reasoning, objective evidence priority, conflict weakening, avoidance of unnecessary examinations, and treatment safety. This synthesis step abstracts the evaluated trajectory into a future-oriented policy object that guides when to examine, how to interpret evidence, when to diagnose, and how to treat similar patient states.

3.4 Skill Bank Maintenance

The inspector maintains the ClinicalSkill Bank by first converting a draft skill card into a mature skill and then fusing it with compatible skills in the bank. Given the draft card $\hat{s}$ produced by the controller, the inspector applies a rule set $\mathcal{R}_{\text{insp}}$ to

perform structural checking, safety checking, and leakage auditing according to predefined criteria:

$$s^\star = \mathcal{M}_{\text{insp}}\left(\hat{s}, \mathcal{R}_{\text{insp}}\right), \tag{10}$$

where $s^\star$ denotes the mature ClinicalSkill card after inspection. A draft card is rejected if required modules are missing, diagnosis–treatment consistency is violated, safety constraints are absent, or patient-specific identifiers, raw trajectory snippets, or case-specific reference answers are retained. Otherwise, the inspector normalizes terminology, removes redundant descriptions, and preserves stricter safety constraints.

To prevent unsafe fusion across clinically distinct skills, the inspector first restricts comparison to a compatible subset of the skill bank:

$$\mathcal{C}(s^\star) = \{\, s_j \in \mathcal{B} \mid \kappa(s_j, s^\star) = 1 \,\}. \tag{11}$$

Here, $\kappa(s_j, s^\star) = 1$ indicates that s_j and $s^\star$ share compatible outcome labels, correctness labels, and safety constraints.

Within the compatible subset, the inspector computes module-level similarity between the mature card $s^\star$ and an existing skill s_j:

$$\text{Sim}(s^\star, s_j) = \sum_m \lambda_m \, \text{sim}(m^\star, m_j), \tag{12}$$

where the four modules correspond to trigger, examination, diagnosis, and treatment, respectively, $\lambda_t + \lambda_e + \lambda_r + \lambda_u = 1$. The best matched skill is selected as:

$$s^\dagger = \arg\max_{s_j \in \mathcal{C}(s^\star)} \text{Sim}(s^\star, s_j), \tag{13}$$

$$\rho^\star = \max_{s_j \in \mathcal{C}(s^\star)} \text{Sim}(s^\star, s_j). \tag{14}$$

If $\rho^\star$ exceeds the merge threshold ρ, the inspector fuses $s^\star$ with the matched skill $s^\dagger$; otherwise, $s^\star$ is inserted as a new skill:

$$\tilde{s} = \text{Fuse}\left(s^\dagger, s^\star; \mathcal{R}_{\text{merge}}\right), \tag{15}$$

$$\mathcal{B}^+ = \begin{cases} (\mathcal{B} \setminus \{s^\dagger\}) \cup \{\tilde{s}\}, & \rho^\star \geq \rho, \\ \mathcal{B} \cup \{s^\star\}, & \rho^\star < \rho. \end{cases} \tag{16}$$

The fusion operation performs semantic union rather than raw concatenation: reusable clinical cues are preserved, duplicate descriptions are removed, and stricter safety constraints are retained. Unlike retrieved case memories, fused ClinicalSkills do not preserve patient-specific trajectories; they preserve reusable action constraints and decision rules for future patient states.

The updated bank $\mathcal{B}^+$ is reused in subsequent episodes through skill retrieval. Overall, the maintenance process closes the self-evolution loop:

$$\tau \rightarrow (\epsilon, \tilde{\tau}) \rightarrow \hat{s} \rightarrow s^\star \rightarrow \mathcal{B}^+. \tag{17}$$

Thus, completed clinical interactions are continuously evaluated, distilled, inspected, and fused into a maintainable ClinicalSkill Bank.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate ClinicalSkill on the MIMIC-CDM benchmark derived from MIMIC-IV, a large-scale de-identified electronic health record database (Johnson et al., 2023; Goldberger et al., 2000). MIMIC-CDM provides an interactive clinical decision-making setting where agents sequentially request examinations, observe returned evidence, update patient states, and make final clinical decisions (Hager et al., 2024a,b). Following the original setting, we evaluate four abdominal diseases: appendicitis, cholecystitis, diverticulitis, and pancreatitis. We use 2,000 cases for self-evolution to construct the ClinicalSkill bank and evaluate on 400 held-out cases. Each test episode requires multi-turn evidence acquisition, final diagnosis, and treatment recommendation.

The self-evolution cases and evaluation cases are strictly disjoint to avoid evaluation leakage. Reference diagnoses and treatment information are used only for post-episode evaluation during skill construction on the self-evolution split; they are never exposed to the agent during held-out testing. During inference, the agent can only access the current patient state, returned examination results, and the maintained ClinicalSkill Bank.

Models and Baselines. We evaluate four instruction tuned open weight LLMs: Meta-Llama-3.2-3B-Instruct (Meta AI, 2024), Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct, and Qwen2.5-32B-Instruct (Yang et al., 2024). We compare ClinicalSkill with vanilla LLM agents and DxEvolve-enhanced agents. DxEvolve is a strong self-evolving diagnostic baseline that performs workflow-aligned evidence acquisition and reuses diagnostic cognition primitives distilled from prior cases (Ren et al., 2026). ClinicalSkill uses the same interaction environment but replaces case-level experience reuse with

Table 1: Main results on the MIMIC-CDM diagnosis-treatment benchmark. We report Diagnostic Accuracy (DA) and Treatment Accuracy (TA) across four diseases and overall. ClinicalSkill uses Qwen2.5-32B-Instruct as the backbone and achieves the best overall performance. Dark blue and light blue backgrounds indicate the best and second-best results, respectively.

Methods	**Appendicitis**		**Cholecystitis**		**Diverticulitis**		**Pancreatitis**		**Overall**	
	DA	**TA**	**DA**	**TA**	**DA**	**TA**	**DA**	**TA**	**DA**	**TA**
General Large Language Models										
Meta-Llama-3.2-3B-Instruct	8.30	73.80	14.01	68.03	6.50	47.39	9.01	38.37	9.82	62.60
Meta-Llama-3-8B-Instruct	54.74	68.94	18.15	65.88	17.36	51.99	11.59	47.93	32.65	62.28
Qwen2.5-7B-Instruct	10.75	26.75	20.83	41.36	10.28	30.84	17.26	41.18	14.76	34.02
Qwen2.5-32B-Instruct	84.97	71.28	66.16	73.54	46.25	51.92	41.71	68.53	67.54	69.58
Self-Evolving Diagnostic Baseline DxEvolve										
Meta-Llama-3.2-3B-Instruct	8.36	72.24	14.07	68.58	6.53	48.44	8.23	40.41	9.71	62.58
Meta-Llama-3-8B-Instruct	75.60	70.95	31.53	67.06	28.19	58.46	24.15	58.20	48.89	66.16
Qwen2.5-7B-Instruct	7.36	28.22	21.94	48.65	9.03	39.46	21.10	40.16	14.26	37.20
Qwen2.5-32B-Instruct	87.50	68.05	82.17	74.75	54.58	46.82	46.83	71.61	74.76	69.43
ClinicalSkill (Ours)	**92.70**	**74.99**	**83.56**	**80.31**	**63.06**	**59.99**	**79.70**	**74.83**	**84.90**	**75.04**

explicit clinical action policies maintained in a skill bank.

Metrics. We report Diagnostic Accuracy (DA) and Treatment Accuracy (TA) as primary metrics. DA is computed as the case-level correctness of the final diagnosis after label normalization. TA is computed by rule-based comparison between the generated treatment plan and the reference treatment information, considering core treatment actions, supportive care, and contraindication-related constraints. For process quality, we compute laboratory F1 and imaging F1 by comparing the requested examinations with canonicalized reference examination sets. All metrics are computed per case and then averaged over the held-out evaluation set. Additional metric details are provided in Appendix A.3.

4.2 Main Results

Table 1 reports the main diagnosis-treatment results. ClinicalSkill achieves the best overall performance, reaching 84.90 % DA and 75.04% TA. Compared with the strongest vanilla LLM baseline, ClinicalSkill improves DA by 17.36 % and TA by 5.46 %. It also outperforms the strongest DxEvolve-enhanced baseline, showing that explicit clinical action policies provide gains beyond case-level diagnostic experience reuse. TA is reported as an independent treatment-quality metric, while the stricter joint diagnosis-treatment correctness is separately analyzed in Figure 5. Paired case-level significance tests further confirm that the main improvements are statistically reliable (Appendix A.5).

The improvement is consistent across all four diseases. ClinicalSkill obtains the highest diagnostic accuracy on appendicitis, cholecystitis, diverticulitis, and pancreatitis, and also improves treatment accuracy overall. The gains are especially clear on more challenging categories such as diverticulitis and pancreatitis, where the agent must coordinate examination selection, diagnostic commitment, and treatment planning under uncertainty. These results support our hypothesis that distilling complete diagnosis-treatment trajectories into explicit clinical action policies yields a more transferable form of medical self-evolution than retrieving retrospective experience alone.

4.3 Process Quality and Difficulty-Stratified Analysis

We next examine whether ClinicalSkill improves the clinical decision process itself during interaction. Figure 3 compares laboratory and imaging F1 between DxEvolve and ClinicalSkill. On Qwen2.5-32B, ClinicalSkill improves laboratory F1 from 84.15% to 87.96% and imaging F1 from 59.98% to 68.72%. The larger gain on imaging suggests that ClinicalSkill is particularly useful for higher-level evidence acquisition decisions, where the agent must identify which examination is clinically most informative rather than indiscriminately requesting more information.

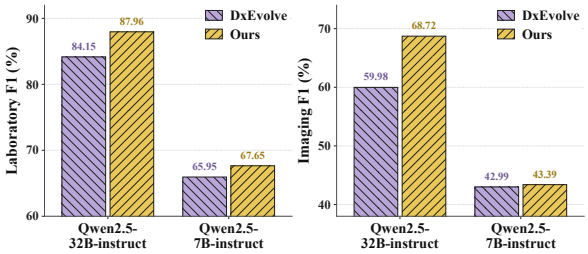

Figure 3: Intermediate process quality. We compare laboratory F1 and imaging F1 between DxEvolve and ClinicalSkill on Qwen2.5-32B-instruct and Qwen2.5-7B-instruct. ClinicalSkill improves both laboratory and imaging alignment, showing that explicit clinical action policies guide better evidence acquisition rather than only improving final answers.

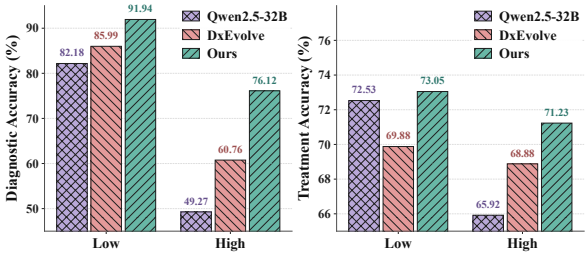

Figure 4: Difficulty-stratified performance. We compare Qwen2.5-32B, DxEvolve, and Ours on low- and high-difficulty cases. ClinicalSkill consistently improves both Diagnostic Accuracy and Treatment Accuracy, with particularly large diagnostic gains on high-difficulty cases.

Figure 4 further stratifies cases by difficulty. ClinicalSkill consistently outperforms both Qwen2.5-32B and DxEvolve under low- and high-difficulty settings. The most notable gain appears in high-difficulty diagnosis, where ClinicalSkill improves over DxEvolve from 60.76% to 76.12%. This indicates that explicit clinical action policies are especially helpful when case-level retrieval is insufficient and the agent must follow a more structured diagnostic path.

4.4 Efficiency and Early-Decision Quality

We further evaluate whether ClinicalSkill improves interaction efficiency. Table 2 shows that ClinicalSkill requires the fewest average rounds, reducing the interaction length from 6.73 rounds for the Qwen2.5-32B to 4.94 rounds, corresponding to a 26.60% reduction. This suggests that ClinicalSkill does not improve performance by requesting more evidence, but by guiding the agent toward more targeted clinical actions.

Figure 5 analyzes early-decision quality and long-tail behavior. ClinicalSkill achieves the highest strict diagnostic accuracy within both 5 and 6 turns, reaching 71.25% and 79.00%, respectively. It also performs best under the stricter joint metric requiring both diagnosis and treatment to be correct. Meanwhile, the proportion of cases requiring at least 10 turns drops from 13.00% for the Qwen2.5-32B to 2.75%. These results show that ClinicalSkill helps agents reach correct diagnosis-treatment decisions earlier while avoiding inefficient long-tail interactions.

Table 2: Interaction efficiency. We report the average number of rounds required to complete multi-turn diagnosis-treatment interaction. Dark blue and light blue backgrounds denote the best and second-best performance, respectively. The red downward arrow indicates the reduction in average interaction rounds compared to the baseline.

Metrics	Methods		
	LLM	DxEvolve	Ours
Rounds	6.73	5.15↓1.58	**4.94**↓1.79
Decrease	/	23.48%	**26.60%**

Table 3: Ablation study on MIMIC-CDM. We report the overall Diagnostic Accuracy (DA) and Treatment Accuracy (TA) by removing each ClinicalSkill module. Dark blue and light blue shadings highlight the best and second-best results, respectively, while the blue upward arrow denotes the gain compared to the baseline.

Methods	Metrics	
	DA	**TA**
ClinicalSkill	**84.90**↑17.36	**75.04**↑5.46
w/o Examination policy	72.50↑4.96	70.86↑1.28
w/o Diagnosis decision rule	73.63↑6.09	70.30↑0.72
w/o Treatment policy	82.09↑14.55	72.21↑2.63

4.5 Ablation Study

We conduct ablation studies to understand which components make a ClinicalSkill an executable action policy rather than a generic case summary. As shown in Table 3, removing the examination policy causes the largest drop in DA, indicating that structured evidence acquisition is the key driver of multi-turn clinical reasoning. Removing the diagnostic decision rule also substantially degrades performance, showing that explicit evidence interpretation and diagnostic commitment criteria are necessary for reliable diagnosis.

The treatment policy mainly affects treatment-specific decision-making. Without this module,

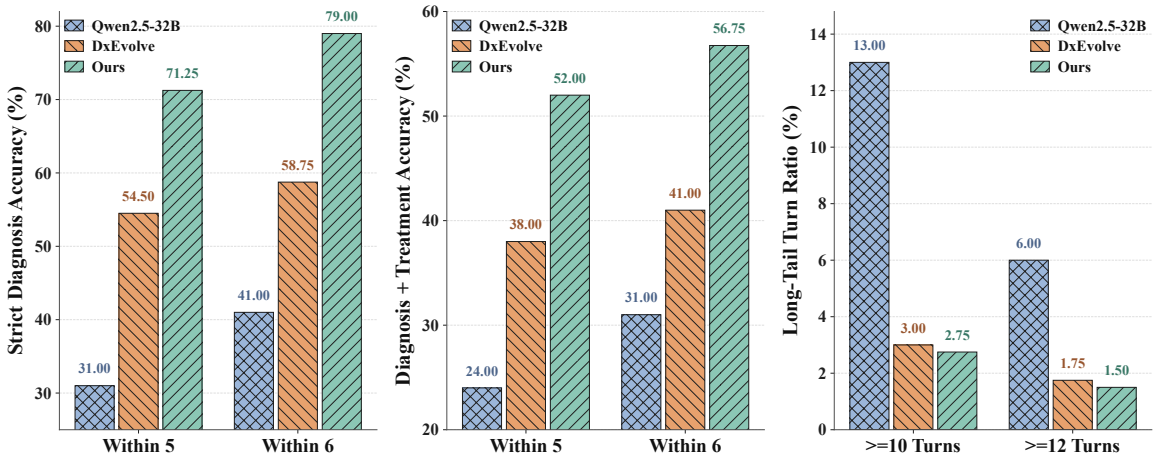

Figure 5: Early-decision quality and long-tail efficiency. We compare strict diagnosis accuracy within 5 and 6 turns, joint diagnosis-treatment accuracy within 5 and 6 turns, and long-tail turn ratios. ClinicalSkill achieves higher early accuracy while reducing long-tail interactions, showing that skill-guided policies improve both correctness and efficiency.

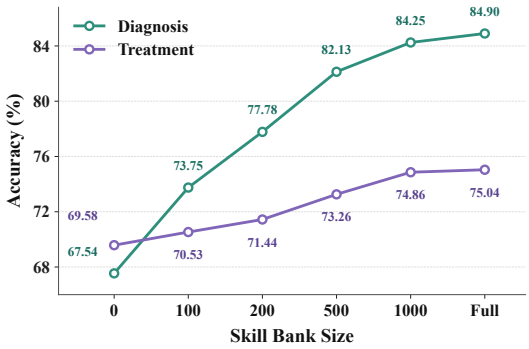

Figure 6: Effect of skill bank size. Diagnostic and treatment accuracy improve as the skill bank grows, with rapid gains at early stages and diminishing returns after sufficient skill coverage.

Table 4: Sensitivity to retrieval size K. Dark blue and light blue shadings highlight the best and second-best results, respectively, while the blue upward arrow denotes the gain compared to the baseline.

Metrics	Top-K		
	1	3	5
DA	81.16↑13.62	**84.90**↑17.36	83.50↑15.96
TA	71.22↑1.64	**75.04**↑5.46	71.70↑2.12

treatment accuracy drops from 75.04% to 72.21%, while diagnostic accuracy remains relatively high. This confirms that treatment planning benefits from diagnosis-aware and safety-aware treatment skills, rather than being a simple by-product of correct diagnosis. Overall, all three modules contribute to ClinicalSkill, and the full model achieves the best performance on both diagnostic and treatment metrics. These ablations show that ClinicalSkill gains come from structured action-policy modules rather than from adding an unconstrained textual memory.

4.6 Parameter Sensitivity

Figure 6 shows that both diagnostic and treatment accuracy improve as the skill bank grows, with rapid gains at early stages and diminishing returns after sufficient coverage. This suggests that ClinicalSkill benefits from accumulating reusable action policies, while redundant skills provide limited additional gains. We further study retrieval size K in Table 4. Top-3 retrieval achieves the best trade-off between sufficient clinical guidance and retrieval noise.

5 Conclusion

We introduced ClinicalSkill, a self-evolving framework distilling diagnosis-treatment trajectories into explicit action policies. By converting completed clinical interactions into reusable policy abstractions, ClinicalSkill guides future decisions with explicit evidence use and safety-aware reasoning. Experiments on MIMIC-CDM demonstrate ClinicalSkill outperforms baselines in diagnosis, treatment, and efficiency. These results show that policy-level self-evolution can improve not only final clinical decisions, but also how agents acquire evidence and decide. Overall, our findings suggest medical agent self-evolution should move beyond storing experience to learning governable policies for future clinical actions.

Limitations

This study has several limitations. First, our evaluation is conducted on MIMIC-CDM derived from MIMIC-IV, which provides a controlled multi-turn diagnosis-treatment setting but may not fully capture the complexity of prospective clinical practice. Second, ClinicalSkill relies on LLM-based evaluator, controller, and inspector modules, so skill quality can still be affected by model errors or imperfect evaluation rules. Third, although our skill bank is designed to be auditable and maintainable, systematic clinician review and real-world safety validation remain necessary before deployment. Future work will extend ClinicalSkill to broader disease domains, prospective clinical workflows, and clinician-in-the-loop skill governance.

References

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Jinrui Fang, Runhan Chen, Xu Yang, Jian Yu, Jiawei Xu, Ashwin Vinod, Wenqi Shi, Tianlong Chen, Heng Ji, ChengXiang Zhai, Ying Ding, and Yuji Zhang. 2026. Benchmarking multi-turn medical diagnosis: Hold, lure, and self-correction. *arXiv preprint arXiv:2604.04325*.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, and 1 others. 2025. A Survey of Self-Evolving Agents: On Path to Artificial Super Intelligence. *arXiv preprint arXiv:2507.21046*.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–e220.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Maximilian Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024a. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622.

Paul Hager, Friederike Jungmann, and Daniel Rueckert. 2024b. MIMIC-IV-Ext Clinical Decision Making: A MIMIC-IV Derived Dataset for Evaluation of Large Language Models on the Task of Clinical Decision Making for Abdominal Pathologies.

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

Guanyu Jiang, Zhaochen Su, Xiaoye Qu, and Yi R. Fung. 2026. XSkill: Continual Learning from Experience and Skills in Multimodal Agents. *arXiv preprint arXiv:2603.12056*.

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 Anthony Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, and Yijun Wang. 2025. Medguide: Benchmarking clinical decision-making in large language models. *arXiv preprint arXiv:2505.11613*.

Qicai Liu, Zhichao Hu, Tao Huang, Yupeng Niu, Xinche Zhang, Shanwu Ma, Chutong Lin, Goh Kim Huat, Hyeokkoo Eric Kwon, Feng Gao, Xianfu Sun, Zhitao Ying, and Guangliang Qiang. 2026. EvoMDT: A Self-Evolving Multi-Agent System for Structured Clinical Decision-Making in Multi-Cancer. *npj Digital Medicine*, 9:124.

Yitao Liu, Chenglei Si, Karthik R. Narasimhan, and Shunyu Yao. 2025. Contextual experience replay for self-improvement of language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 14179–14198.

Meta AI. 2024. Llama 3.2 Model Cards and Prompt Formats. `https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/`. Accessed: 2026-05-13.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P. Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. 2025. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*.

Ruiyang Ren, Yuhao Wang, Yunsen Liang, Lan Luo, Jing Liu, Haifeng Wang, Cong Feng, Yinan Zhang,

Chunyan Miao, Ji-Rong Wen, and Wayne Xin Zhao. 2026. Emulating clinician cognition via self-evolving deep clinical research. *arXiv preprint arXiv:2603.10677*.

Weixiang Shen, Bailiang Jian, Jun Li, Che Liu, Johannes Moll, Xiaobin Hu, Daniel Rueckert, Hongwei Bran Li, and Jiazhen Pan. 2026. Evo-medagent: Beyond one-shot diagnosis with agents that remember, reflect, and improve. *arXiv preprint arXiv:2604.14475*.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, Arman Cohan, and Mark Gerstein. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.

Jiongxiao Wang, Qiaojing Yan, Yawei Wang, Yijun Tian, Soumya Smruti Mishra, Zhichao Xu, Megha Gandhi, Panpan Xu, and Lin Lee Cheong. 2025. Reinforcement learning for self-improving agent with skill library. *arXiv preprint arXiv:2512.17102*.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. 2025. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *Preprint*, arXiv:2511.20857.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2025. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning. *arXiv preprint arXiv:2508.19828*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Lin Yang, Yuancheng Yang, Xu Wang, Changkun Liu, and Haihua Yang. 2026a. Medmt-bench: Can llms memorize and understand long multi-turn conversations in medical scenarios? *arXiv preprint arXiv:2603.23519*.

Yutao Yang, Junsong Li, Qianjun Pan, Bihao Zhan, Yuxuan Cai, Lin Du, Jie Zhou, Kai Chen, Qin Chen, Xin Li, Bo Zhang, and Liang He. 2026b. Autoskill: Experience-driven lifelong learning via skill self-evolution. *arXiv preprint arXiv:2603.01145*.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. 2026. Memskill: Learning and evolving memory skills for self-evolving agents. *arXiv preprint arXiv:2602.02474*.

A Appendix

A.1 ClinicalSkill Card Schema

A ClinicalSkill card represents an explicit clinical action policy rather than a free-form case summary. It contains six modules: trigger, examination, diagnosis, treatment, prompt, and metadata. Together, these modules specify when the skill applies, how evidence should be acquired and interpreted, how treatment should be recommended, and how provenance should be preserved for auditing and maintenance.

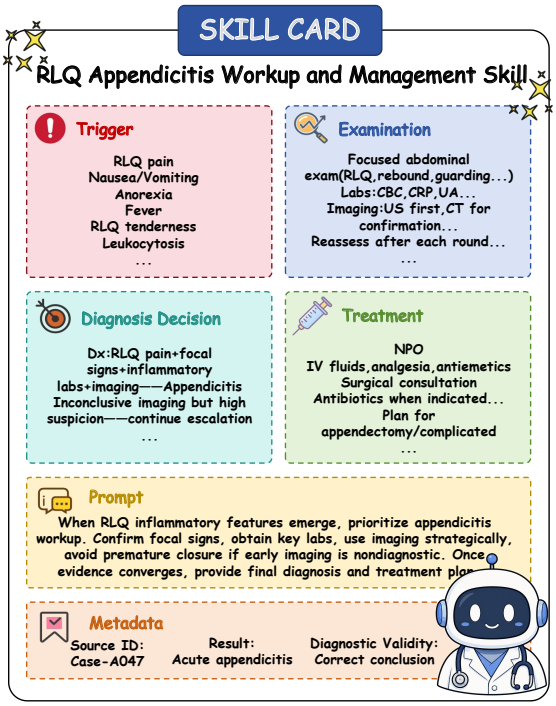

Figure A1: Example ClinicalSkill card. A skill card contains trigger, examination, diagnosis, treatment, prompt, and metadata modules. The card is used as an explicit clinical action policy to guide future multi-turn diagnosis and treatment decisions.

Formally, each skill card follows the schema

$$s_i = (t_i, e_i, r_i, u_i, \pi_i, \mu_i),$$

where t_i is the trigger module, e_i is the examination policy, r_i is the diagnostic decision rule, u_i is the treatment policy, π_i is the compiled runtime prompt, and μ_i stores provenance and maintenance metadata. Table A1 summarizes the role of each module.

Table A1: ClinicalSkill card schema. Each module contributes to an auditable and executable clinical action policy.

Module	Description
Trigger t_i	Defines applicability conditions, including key symptoms, abnormal findings, temporal cues, exclusion signals, and retrieval keywords.
Examination e_i	Specifies evidence acquisition actions, including physical examination, laboratory tests, imaging, escalation conditions, and stop criteria.
Diagnosis r_i	Encodes diagnostic reasoning rules, including supporting evidence, refuting evidence, differential diagnoses, and diagnostic commitment criteria.
Treatment u_i	Provides diagnosis-aware treatment recommendations, including primary treatment, supportive care, escalation rules, contraindication checks, and monitoring suggestions.
Prompt π_i	Compiles the structured modules into a runtime instruction that can be injected into the medical agent.
Metadata μ_i	Records provenance information, including source episode, evaluated diagnosis, treatment label, correctness signal, timestamp, and maintenance history.

A.2 Skill Construction and Maintenance

ClinicalSkill is constructed after a complete multi-turn diagnosis-treatment episode. The agent first interacts with a patient case by requesting examinations, observing returned evidence, updating the patient state, and finally producing a diagnosis and treatment recommendation. The completed trajectory is then processed by three LLM-based modules: evaluator, controller, and inspector.

Evaluator. The evaluator takes the completed trajectory, reference information, and an evaluation rulebook as input. It checks whether examination actions are necessary and properly ordered, whether the final diagnosis is supported by sufficient evidence, whether the treatment recommendation is consistent with the diagnosis, and whether unsafe or unsupported actions appear in the trajectory. The evaluator is used only after an episode is completed for skill construction, and does not intervene in held-out test-time decision making.The evaluator uses reference information only to assess completed self-evolution episodes and does not expose case-specific reference answers to future held-out cases. During testing, retrieved skills contain only abstracted policy mod-

ules rather than original case labels or patient-specific answers.

Controller. The controller converts the evaluated trajectory into a candidate ClinicalSkill card. It does not copy the trajectory verbatim. Instead, it abstracts reusable decision structure into the predefined schema: trigger, examination, diagnosis, treatment, prompt, and metadata. This turns a completed clinical episode into a future-oriented policy object that can guide similar patient states.

Inspector. The inspector maintains the skill bank by validating, clustering, and merging candidate skills. It rejects malformed skills or skills that violate clinical invariants, such as missing safety constraints or inconsistent diagnosis-treatment pairs. Valid skills are compared only with clinically compatible skills that share the same outcome and correctness labels. During merging, the inspector preserves reusable clinical cues, removes redundant descriptions, and retains stricter safety constraints.

A.3 Evaluation Details

All evaluation metrics are computed at the case level and then averaged over the held-out evaluation set. For each test case, the agent first completes a multi-turn interaction, outputs a final diagnosis, and then provides a treatment recommendation. The reference labels and treatment information are used only for evaluation and are never exposed to the agent during testing.

Diagnostic Accuracy. Diagnostic Accuracy (DA) measures whether the final diagnosis matches the reference disease label after normalization. Each case receives a case-level diagnostic score after label normalization, and the reported DA is the mean score across all held-out cases.

Treatment Accuracy. Treatment Accuracy(TA) evaluates whether the generated treatment plan is consistent with the reference treatment information. We use rule-based matching over normalized treatment elements, including core treatment actions, supportive care, escalation or de-escalation decisions, and contraindication related constraints. Each case receives a case-level treatment score, and the reported TA is averaged over the held-out cases.

Process Metrics. Laboratory F1 and Imaging F1 measure whether the agent requests clinically appropriate evidence during the interaction. For each case, requested laboratory and imaging examinations are canonicalized and compared with the corresponding reference examination sets. We compute precision, recall, and F1 at the case level, and report the average F1 across cases.

Efficiency Metrics. Efficiency is evaluated by the average number of interaction rounds, early-decision accuracy within fixed turn budgets, and long-tail turn ratios. These metrics assess whether an agent can reach correct diagnosis–treatment decisions efficiently rather than relying on prolonged interaction.

A.4 Top-K Retrieval Sensitivity

We analyze the effect of the number of retrieved skills during inference. As shown in Table 4, retrieving Top-3 skills achieves the best overall performance. Retrieving only one skill may provide insufficient guidance for ambiguous cases, while retrieving five skills may introduce less relevant information and weaken decision focus. All Top-K variants remain stronger than the strongest baseline, indicating that ClinicalSkill is robust to retrieval-size variation.

A.5 Statistical Analysis

Table A2: Paired statistical significance tests for the main Qwen2.5-32B-Instruct result on the 400-case held-out set. All p-values are from two-sided paired Wilcoxon signed-rank tests.

Comparison	Metric	p-value
LLM vs Ours	DA	1.24×10^{-18}
DxEvolve vs Ours	DA	8.25×10^{-10}
LLM vs Ours	TA	3.96×10^{-8}
DxEvolve vs Ours	TA	8.20×10^{-13}
LLM vs Ours	Rounds	4.28×10^{-41}
DxEvolve vs Ours	Rounds	5.77×10^{-4}

We conduct paired case-level significance tests on the 400 held-out cases to assess whether the main improvements of ClinicalSkill are statistically reliable. Since all methods are evaluated on the same cases, we use a paired design. For DA, TA, and interaction rounds, we compare ClinicalSkill with the vanilla backbone and the strongest DxEvolve-enhanced baseline using two-sided paired Wilcoxon signed-rank tests.Since DA and TA are computed as graded case-level scores

rather than binary outcomes, we use two-sided paired Wilcoxon signed-rank tests for paired significance analysis.

As shown in Table A2, ClinicalSkill significantly improves both primary metrics and interaction efficiency over the compared baselines. These results indicate that the observed gains are consistent across held-out cases rather than being driven by a small number of outliers.