MEDGUIDEEVOLVE: EVOLVING TRANSFERABLE CLINICAL EXPERIENCE CARDS FROM CASES UNDER ESTABLISHED GUIDELINES

Anonymous authors
Paper under double-blind review

ABSTRACT

Diagnosing rare cardiovascular diseases is a complex clinical reasoning task that requires integrating clinical evidence across time and specialties and differentiating among potential etiologies underlying overlapping phenotypes. How to distill guideline-application experience that transfers across patients from multi-case clinical trajectories while remaining faithful to an established clinical guideline remains a key unresolved problem in the field. Direct case reuse relies on inter-patient phenotypic similarity and therefore struggles to determine whether phenotypic similarity implies a transferable rationale for guideline application, while experience abstraction can overfit limited cases by encoding incidental patterns and losing the original evidentiary context and applicability conditions during generalization. We propose **MedGuideEvolve**, which compares erroneous cases, similar correctly diagnosed cases, and hard controls under an established clinical guideline to distill conditional guideline-application experience with explicit evidence links and applicability conditions. Candidate experience bundles are retained only when they satisfy schema, provenance, guideline, applicability, audit, support, and paired-repair gates, including non-negative paired Top-1 net repair on independent patients held out from experience formation. This admission boundary limits overgeneralization from incidental patterns while leaving final diagnostic adjudication to the established clinical guideline. On 217 label-blinded cases from a temporally held-out test set, MedGuideEvolve improves Qwen3.8-27B's Top-1 accuracy over Base by 23.96 percentage points to 82.95%, demonstrating that conditional guideline-application experience distilled within established guideline constraints can transfer to unseen patients while retaining provenance.

1 INTRODUCTION

Diagnosing rare cardiovascular diseases requires integrating evidence across time and specialties while distinguishing etiologies that produce overlapping phenotypes. LLM-based medical agents now combine retrieval, tool use, feedback, and multi-agent coordination to support complex clinical reasoning (Kim et al., 2024; Zhao et al., 2026; Chen et al., 2026). This setting calls for evidence-centered disease reasoning: diseases are sparsely represented, decisive clues span history, electrocardiography, imaging, laboratory tests, genetics, and extracardiac findings, and some clues emerge only over the disease course (Mao et al., 2025). The 2023 ESC cardiomyopathy guideline therefore calls for phenotype-first, multimodal, and systematic etiologic assessment (Arbelo et al., 2023). Retrieval and multi-agent systems support each patient episode, creating an opportunity to accumulate and reuse diagnostic experience across patients (Ou et al., 2026; Li et al., 2024). The central question is how an agent can turn recurring cross-case diagnostic errors into an evidence-linked review operation that improves current-patient reasoning with diagnostic adjudication grounded in the guideline throughout each patient episode.

We organize prior approaches by how historical information enters reasoning. Direct case reuse retrieves concrete precedents from case reports or similar-patient resources and exposes their evidence combinations, temporal relations, or diagnostic processes as context for the current patient (Fujiwara et al., 2018; Alsentzer et al., 2025)(Shi et al., 2024). Abstract reuse distills reflections, rules, diagnostic insights, or procedural strategies from successful or failed trajectories and applies them

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.

to later patients (Shinn et al., 2023; Zhao et al., 2024). More structured memories turn these abstractions into reusable procedures, hierarchical workflows, or skill repositories (Wang et al., 2024; Fan et al., 2026; Sun et al., 2026). Direct reuse provides rich clinical detail, abstract reuse provides transferable patterns, and structured memory provides reusable organization. The design opportunity is to combine these strengths in an experience card: a provenance-linked set of evidence conditions and a context-specific review action whose validated applicability matches the evidence available in the current patient. Figure 1 summarizes this progression: direct case reuse offers concrete case detail, abstract reuse offers transferable patterns, and MedGuideEvolve combines these strengths into reusable evidence-linked review experience that guides future patient-level reasoning.

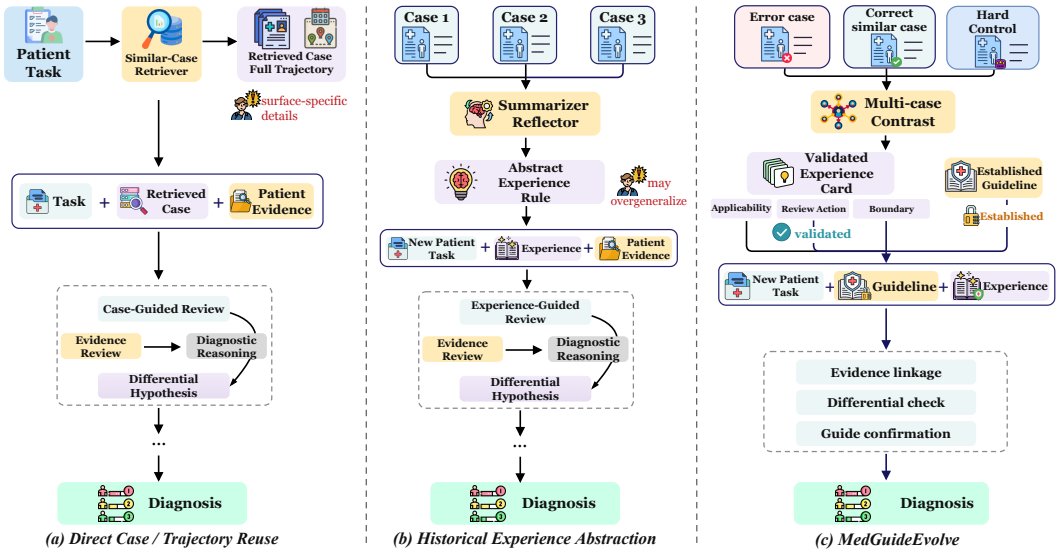


Figure 1: Motivation for evidence-linked experience: case reuse retains detail; abstraction transfers patterns but may overgeneralize; MedGuideEvolve combines both in guideline-linked cards.

The experience card matches the evidence shared with the current patient. A full phenotype-and-course match supports a concrete evidence-checking sequence; a high-level match such as increased wall thickness activates a broader review strategy: *look again for corroborating and differentiating evidence*. Wall thickening becomes a useful review trigger when voltage–wall-thickness discordance beyond the pattern expected from hypertension is accompanied by infiltrative imaging or extracardiac manifestations; when hypertension adequately explains the finding, the guideline's direct pathway provides the appropriate route (Arbelo et al., 2023). Together, the Guide Library defines evidence obligations, the Case Bank shows how these rules apply to concrete patients, and the Experience Library routes the appropriate strategy for solving new patient scenarios.

MedGuideEvolve turns these cross-case contrasts into guideline-linked experience cards. Its Snapshot–Induce–Admit–Route cycle incorporates four operations: contrast identifies recurring errors through erroneous cases, similar correctly handled cases, and hard controls; projection aligns candidate skills with guideline evidence obligations; validation tests incremental ranking repair on patient-disjoint paired probes with a non-negative paired Top-1 protection criterion; and routing selects the most specific skill supported by current evidence and combines it with guideline reasoning. This cycle enables the agent to organize evidence more effectively while applying established disease definitions and confirmation criteria throughout each patient-level decision. We evaluate MedGuideEvolve in a label-blinded diagnostic task on the 217 Test cases from the chronological split. On Qwen3.8-27B, it improves Top-1 accuracy over Base by 23.96 percentage points to 82.95%, demonstrating forward transfer to unseen patients.

Our contributions cover experience learning, guideline validation, routing, and evaluation:

- We formulate guideline-anchored experience learning through provenance-linked cards that connect evidence conditions with context-specific review actions.

- We introduce MedGuideEvolve, which contrasts erroneous cases, similar correctly handled cases, and hard controls to construct evidence-matched cards and route them through the guideline for evidence-centered clinical reasoning.
- We develop guideline projection and validation-gated library evolution, validating candidate cards through patient-disjoint paired probes before library updates.
- We establish a patient-disjoint temporal protocol and demonstrate a 23.96-point Top-1 gain over Base for Qwen3.8-27B, reaching 82.95% on the 217 Test cases from the chronological split used for forward-transfer evaluation of unseen patients.

2 Related Work

LLM-based Medical Agents. Medical agents extend clinical reasoning with retrieval, specialty tools, and multi-agent coordination. MDAgents adapts collaboration to task complexity, while DeepRare and RareAgents combine coordinated agents with traceable resources for rare-disease diagnosis (Kim et al., 2024; Zhao et al., 2026; Chen et al., 2026). Case-based methods address sparse disease coverage through phenotype matching and similar-patient retrieval, as in PubCaseFinder and SHEPHERD (Fujiwara et al., 2018; Alsentzer et al., 2025). Other systems rank standardized phenotypes or retrieve structured EHR experience, including PhenoBrain and EXPRAG (Mao et al., 2025; Ou et al., 2026). Despite these capabilities, historical cases generally remain episode-level context, leaving the transferable evidence relations and applicability boundaries under-specified. MedGuideEvolve converts historical cases into provenance-linked conditional review experience, while the established guideline remains authoritative over diagnostic state during cross-patient transfer.

Self-Evolving Memory. Experience reuse in LLM agents ranges from storing trajectories and reflections to distilling reusable workflows and skills. Early methods preserve interaction traces or verbal feedback (Shinn et al., 2023; Zhang et al., 2023), whereas later approaches extract reusable procedures and memories (Zhao et al., 2024; Wang et al., 2024; Xu et al., 2026). MetaFlowLLM represents experience as hierarchical workflows, while SkeMex distills medical trajectories into structured skills and uses context-dependent utility for retrieval and repository governance (Fan et al., 2026; Sun et al., 2026). These mechanisms make experience reusable, but they do not make a guideline-defined evidence obligation the admission boundary for cross-patient transfer: trajectory memories can retain instance-specific details, while abstract skills can omit triggers, provenance, and failure boundaries. A clinical guideline such as the 2023 ESC cardiomyopathy guideline specifies diagnostic boundaries and confirmation criteria, but not reusable cross-case evidence-review operations (Arbelo et al., 2023). MedGuideEvolve couples experience reuse with explicit evidence conditions and patient-disjoint validation, then falls back to the established guideline when evidence match is insufficient for transfer across patients and diseases.

3 Method

MedGuideEvolve follows a Snapshot–Induce–Admit–Route cycle for learning transferable operations for applying an established guideline across patients (Figure 2). The guideline graph $\mathcal{G}$ defines evidence obligations, legal transitions, and diagnostic guards; the Case Bank $\mathbb{C}$ stores locked trajectories for offline contrast, and the Experience Library $\mathbb{M}_t$ stores validated review strategies. Offline, the cycle freezes evidence, induces and projects a candidate bundle, and commits it only after patient-disjoint admission checks; online, the resulting library is frozen and routes evidence-matched experience through guideline-constrained diagnosis. Section 3.1 defines the snapshot, application contract, and Experience Skill; Section 3.2 formulates contrastive induction; Section 3.3 validates updates through paired probes; and Section 3.4 presents hierarchical routing and online diagnosis, with the library organizing evidence review and the established guideline adjudicating diagnostic state. The corresponding offline and online execution order is given in Algorithms 1 and 2 in Appendix C.

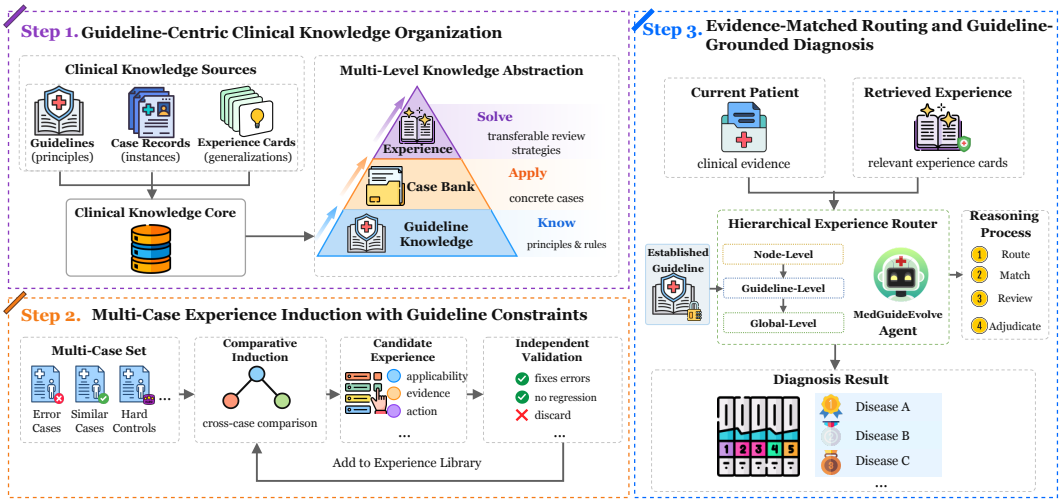


Figure 2: Overview of MedGuideEvolve. Offline multi-case induction proposes candidate experience cards, and independent patient-disjoint validation gates library updates; at deployment, the established guideline adjudicates diagnosis while evidence-matched cards guide evidence review.

3.1 Guideline-Constrained State and Experience Objects

Formally, MedGuideEvolve learns the Experience Library $\mathbb{M}^\star$ while the established-guideline graph $\mathcal{G}$, model parameters $\boldsymbol{\theta}$, and generation contract Γ remain fixed. The locked Case Bank $\mathbb{C}$ stores trajectories, while at round t, $\mathbb{M}_t \subseteq \mathbb{S}$ is the current library and $\mathbb{B}_t \subseteq \mathbb{S}$ is a candidate bundle; $\mathbb{S}$ denotes the Experience Skill universe. Patient snapshots satisfy $\boldsymbol{x}_p \in \mathbb{X}$ and path/node states satisfy $\boldsymbol{s} \in \mathbb{H}$, with $\mathbb{X}$ and $\mathbb{H}$ denoting the corresponding spaces. We use p, t, i, q, ω, and j for a patient, round, probe, evidence item, predicate, and node step, respectively; the five-valued assertion-state set $\mathbb{V}_{\text{obs}}$ preserves distinct evidence conditions throughout adjudication, validation, and deployment.

For patient p, the executor forms an immutable pre-anchor snapshot from visible messages before diagnosis, recording each observed concept, assertion state, and source position:

$$\boldsymbol{x}_p = ((f_{pq}, v_{pq}, \pi_{pq}))_{q=1}^{m_p}. \tag{1}$$

Here f_{pq} is a clinical concept, $v_{pq} \in \mathbb{V}_{\text{obs}}$ its assertion state, π_{pq} its source position, and m_p the number of visible concepts. The five states in $\mathbb{V}_{\text{obs}}$—*present*, *absent*, *not–done*, *not–recorded*, and *conflicting*—remain distinct during adjudication, and the pre-anchor snapshot is reused across validation arms and retained in the append-only audit trace for reproducible adjudication.

For any experience set $\mathbb{U} \subseteq \mathbb{S}$, we write $\mathcal{A}(\cdot; \boldsymbol{\theta}, \Gamma, \mathcal{G}, \mathbb{U})$ for the contract-defined application map that combines model generation with guideline-constrained adjudication and returns a ranking, adjudicated node states, and an append-only audit trace for a given snapshot and experience set:

$$\left(\widehat{\boldsymbol{y}}_p, \boldsymbol{s}_p, \boldsymbol{z}_p\right) = \mathcal{A}(\boldsymbol{x}_p; \boldsymbol{\theta}, \Gamma, \mathcal{G}, \mathbb{U}). \tag{2}$$

The output ranking $\widehat{\boldsymbol{y}}_p$ has the contract-defined cardinality, while $\boldsymbol{s}_p$ records states reached on active guideline nodes and $\boldsymbol{z}_p$ concatenates per-transition audit events $\delta\boldsymbol{z}_j$. Guideline adjudication accompanies the ranking by assigning confirmation, exclusion, or unresolved status; probe reference outcomes are revealed only after the prediction, trace, and sidecars are locked for evaluation.

A completed trajectory is recorded in the Case Bank as evidence, node, model, outcome, and provenance fields. Its reusable abstraction is a conditional Experience Skill:

$$e = \left(\ell_e, \rho_e, \sigma_e, \mathbb{Q}_e^+, \mathbb{Q}_e^-, a_e, \nu_e\right). \tag{3}$$

Here ℓ_e, ρ_e, and σ_e specify location, role, and validated scope; $\mathbb{Q}_e^+$ and $\mathbb{Q}_e^-$ are required and exclusion predicates; a_e is the scoped review action; and ν_e stores version and provenance, from which $\text{ver}(e)$ is read. At runtime, the card reconnects evidence, compares explanations, or prompts re-evaluation; the established guideline adjudicates the resulting diagnostic node.

3.2 CONTRASTIVE INDUCTION UNDER AN ESTABLISHED GUIDELINE

Offline induction starts from a recurring guideline-application problem. Once the relevant trajectories are locked, their outcomes, evidence, and provenance localize the problem as $\lambda_t = (\ell_t, \kappa_t, \chi_t)$, where ℓ_t is the guideline or node location, κ_t is the error type, and χ_t is the visible evidence context. The induction procedure forms an ordered contrast $\mathbb{K}_t = (\mathbb{K}_t^-, \mathbb{K}_t^+, \mathbb{K}_t^{\mathrm{hard}})$: failure cases expose recurring mistakes, similar correct cases preserve the desired decision under comparable evidence, and hard controls share salient evidence while marking a different direction or boundary.

The locked Case Bank provides the localized problem and its ordered contrast. The proposer combines this contrast with $\mathbb{M}_t$ to construct an unvalidated bundle for arm C; the established guideline projection filters and normalizes the proposal before forming the candidate bundle:

$$\mathbb{B}_t = \Pi_{\mathcal{G}}(\phi(\lambda_t, \mathbb{K}_t, \mathbb{M}_t; \boldsymbol{\theta}, \Gamma)) \,. \tag{4}$$

Here $\phi(\cdot; \boldsymbol{\theta}, \Gamma)$ maps the contrast and $\mathbb{M}_t$ to a finite set of scoped skills with evidence predicates, review actions, and provenance; $\Pi_{\mathcal{G}} : 2^{\mathbb{S}} \to 2^{\mathbb{S}}$ filters and normalizes these fields under the established guideline, producing a candidate bundle $\mathbb{B}_t$ for validation and admission to the library.

3.3 VALIDATION-GATED LIBRARY EVOLUTION

Validation measures the incremental value of the candidate bundle under identical execution conditions and the established guideline. The sealed bundle is evaluated on fresh, patient-disjoint probes; all arms share the normalized snapshot, model, generation contract, and budget, differing only in experience set. For arm $h \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$, let $\mathbb{U}_{\mathrm{A}} = \varnothing$, $\mathbb{U}_{\mathrm{B}} = \mathbb{M}_t$, and $\mathbb{U}_{\mathrm{C}} = \mathrm{apply}(\mathbb{M}_t, \mathbb{B}_t)$:

$$\left(\widehat{\boldsymbol{y}}_i^{\,h}, \boldsymbol{s}_i^{\,h}, \boldsymbol{z}_i^{\,h}\right) = \mathcal{A}(\boldsymbol{x}_i; \boldsymbol{\theta}, \Gamma, \mathcal{G}, \mathbb{U}_h), \qquad h \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}. \tag{5}$$

Arm A is the no-experience reference, B the current library, and C the proposed update; C–B supplies admission evidence, whereas C–A measures cumulative gain. Predictions and sidecars are locked before probe outcomes and independent applicability references are revealed; each batch yields one admission decision, while apply constructs the sealed C library that Eq. (8) commits only when every admission check passes for the candidate bundle.

Before injection, the executor evaluates applicability by matching card location and role, version, and evidence predicates. Let $\boldsymbol{s}$ denote the current guide–node state, let $\mathrm{match}(e, \boldsymbol{s})$ check the card's location and role, and let $\mathrm{eval}(\omega \mid \boldsymbol{x}, \boldsymbol{s})$ return *true*, *false*, or *unknown* for predicate ω; we use $\mathbf{1}_{\{\cdot\}}$ for the Boolean indicator. The applicability indicator is:

$$\begin{aligned}\mathrm{use}_{\mathbb{W}_t^{\mathrm{allow}}}(e \mid \boldsymbol{x}, \boldsymbol{s}) = \mathbf{1}_{\{\mathrm{match}(e,\boldsymbol{s})=1\}} \mathbf{1}_{\{\mathrm{ver}(e) \in \mathbb{W}_t^{\mathrm{allow}}\}} \\ \times \prod_{\omega \in \mathbb{Q}_e^+} \mathbf{1}_{\{\mathrm{eval}(\omega \mid \boldsymbol{x}, \boldsymbol{s}) = \mathrm{true}\}} \prod_{\omega \in \mathbb{Q}_e^-} \mathbf{1}_{\{\mathrm{eval}(\omega \mid \boldsymbol{x}, \boldsymbol{s}) = \mathrm{false}\}} .\end{aligned} \tag{6}$$

Here $\mathbb{W}_t^{\mathrm{allow}}$ contains the versions admitted for the current arm, including the candidate in C. More generally, $\mathrm{use}_{\mathbb{W}}(e \mid \boldsymbol{x}, \boldsymbol{s}) \in \{0, 1\}$ is the applicability predicate under admitted set $\mathbb{W}$; unknown predicate values remain explicit in the audit trace but evaluate to zero, preserving separate recall, applicability, and injection outcomes across validation arms.

For a prespecified Top-k endpoint, let $y_i^\star$ be the reference disease and define $b_{ik} = \mathbf{1}_{\{y_i^\star \in \mathrm{Top}_k(\widehat{\boldsymbol{y}}_i^{\,\mathrm{B}})\}}$ and $c_{ik} = \mathbf{1}_{\{y_i^\star \in \mathrm{Top}_k(\widehat{\boldsymbol{y}}_i^{\,\mathrm{C}})\}}$ for patient i under B and C, respectively. Let N be the paired-probe sample size. The paired net-repair statistic for candidate bundle $\mathbb{B}_t$ is:

$$\mathrm{Net}_k(\mathbb{B}_t) = \sum_{i=1}^{N} \left[\mathbf{1}_{\{b_{ik}=0 \wedge c_{ik}=1\}} - \mathbf{1}_{\{b_{ik}=1 \wedge c_{ik}=0\}}\right] . \tag{7}$$

The statistic compares C with B for bundle $\mathbb{B}_t$; Top-3 paired net repair is the primary admission endpoint, Top-1 is the protection endpoint, and $\mathrm{Net}_k(\mathbb{B}_t)/N$ is the normalized gain. Admission also checks each active card e: let $\mathbb{D}_t(e)$ be its audited reference-unit set, $U_t(e)$ the unknown count, and $\mathbb{P}_t^{\mathrm{inj}}(e)$ its injected probes, with $u_t(e) = U_t(e)/|\mathbb{D}_t(e)|$ and a zero denominator treated as failure. We require $|\mathbb{P}_t^{\mathrm{inj}}(e)| \geq 3$ and $u_t(e) \leq 0.20$ for every active card, together with $\mathrm{Net}_3(\mathbb{B}_t) \geq 1$ and $\mathrm{Net}_1(\mathbb{B}_t) \geq 0$. We set $\mathrm{admit}(\mathbb{B}_t) = 1$ only when schema, provenance, privacy, guideline

consistency, applicability, support, audit integrity, and quantitative checks pass; when they do, the bundle-level rule sets the next library state as follows:

$$\mathbb{M}_{t+1} = \begin{cases} \mathrm{apply}(\mathbb{M}_t, \mathbb{B}_t), & \mathrm{admit}(\mathbb{B}_t) = 1, \\ \mathbb{M}_t, & \mathrm{admit}(\mathbb{B}_t) = 0. \end{cases} \tag{8}$$

An accepted round records the bundle provenance and advances $\mathbb{M}_t$, whereas an unsuccessful round preserves the current library. Validation and commit events remain append-only, and each subsequent round consumes a fresh patient-disjoint probe batch for the next admission decision.

3.4 Hierarchical Routing and Online Clinical Diagnosis

After the final development round, repeated application of Eq. (8) yields the admitted library $\mathbb{M}^\star$, which is held fixed for online execution. Let $\mathbb{W}^\star = \{\mathrm{ver}(e) : e \in \mathbb{M}^\star\}$ denote the versions admitted at deployment. The router activates at most three evidence-backed guideline paths for each new snapshot, queues remaining supported directions, and promotes them as active paths terminate. Routing determines where experience enters, while the established guideline supplies the evidence obligations, entry conditions, and diagnostic transitions for online diagnosis.

Within an active path, let $\boldsymbol{s}$ encode its guideline path and diagnostic node, with $g(\boldsymbol{s})$ denoting the associated guideline. Partition the frozen library by validated scope into the node-, guideline-, and global-scope subsets $\mathbb{M}^{\star,\mathrm{node}}(\boldsymbol{s})$, $\mathbb{M}^{\star,\mathrm{guide}}(g(\boldsymbol{s}))$, and $\mathbb{M}^{\star,\mathrm{global}}$. The selector searches from the most specific validated scope to broader scopes:

$$\mathbb{R}^\star(\boldsymbol{s}) = \mathbb{M}^{\star,\mathrm{node}}(\boldsymbol{s}) \cup \mathbb{M}^{\star,\mathrm{guide}}(g(\boldsymbol{s})) \cup \mathbb{M}^{\star,\mathrm{global}}. \tag{9}$$

Let $\mathbb{I}^\star(\boldsymbol{x}, \boldsymbol{s}) = \{e \in \mathbb{R}^\star(\boldsymbol{s}) : \mathrm{use}_{\mathbb{W}^\star}(e \mid \boldsymbol{x}, \boldsymbol{s}) = 1\}$ be the admissible subset. The selector injects at most one admissible card per node, choosing it by the lexicographic score $r(e \mid \boldsymbol{x}, \boldsymbol{s})$:

$$e^\star(\boldsymbol{x}, \boldsymbol{s}) = \operatorname*{arg\,max}_{e \in \mathbb{I}^\star(\boldsymbol{x}, \boldsymbol{s})} r(e \mid \boldsymbol{x}, \boldsymbol{s}). \tag{10}$$

With the convention $e^\star(\boldsymbol{x}, \boldsymbol{s}) = \perp$ when $\mathbb{I}^\star(\boldsymbol{x}, \boldsymbol{s}) = \varnothing$, the selector uses the lexicographic key $(\mathrm{scope}, \mathrm{overlap}, \mathrm{support}, -\mathrm{order})$: node-specific scope precedes guideline-level scope, which precedes global scope; within each scope, trigger-concept overlap, validated support, and stable card order resolve ties. Similarity retrieves candidates but never grants eligibility; validated scope and evidence predicates determine injection. Each card is evaluated at its recorded disease and node granularity, while global cards retain their validated target-node scope; when $e^\star = \perp$, the node proceeds under the established guideline without injected experience.

At node n_j, let $\boldsymbol{x}$ denote the current patient snapshot and set $e^\star = e^\star(\boldsymbol{x}, \boldsymbol{s}_j)$, with $e^\star = \perp$ when no card is selected. The model proposes evidence references, rule checks, a status, and a successor; the executor recomputes evidence states from the immutable snapshot and applies guideline-constrained adjudication to produce the next path state and audit event:

$$\left(\boldsymbol{s}_{j+1}, n_{j+1}, \delta \boldsymbol{z}_{j+1}\right) = \mathcal{T}_{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{s}_j, n_j, e^\star; \boldsymbol{\theta}, \Gamma)\,. \tag{11}$$

Here $\mathcal{T}_{\mathcal{G}}$ adjudicates each proposal under the established guideline and appends the audit event $\delta \boldsymbol{z}_{j+1}$ to $\boldsymbol{z}_p$ in Eq. (2). The validator preserves the five evidence states and applies guard-supported confirmation or exclusion at each transition. Paths rotate round-robin; a terminal or blocked state, or a repeated state fingerprint, stops only the current path. Execution ends when no supported paths remain or the node budget is exhausted. Finally, a reserved call maps the initial ranking and adjudicated node states to the output schema; MedGuideEvolve's admitted library $\mathbb{M}^\star$ and generation contract remain fixed across patients in the patient-disjoint Temporal Test evaluation, where rankings are scored against reference disease labels.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We curated de-identified cardiovascular rare-disease records from a single hospital into the **CardioRareBench** dataset for cardiovascular rare-disease diagnosis. The dataset

Table 1: Main results on the patient-disjoint Temporal Test for the diagnostic task. **Bold** values denote per-metric best results within each backbone panel; underlined values denote runners-up. Changes from Base are computed before rounding and reported in percentage points; blue and red indicate gains and losses.

Method	Source	Candidate ranking								Class-aware metrics			
		Top-1 Acc. (%)		Recall@3 (%)		Recall@5 (%)		MRR@5 (%)		Macro-F1 (%)		Bal. Acc. (%)	
		Acc.	Δ_{Base}	R@3	Δ_{Base}	R@5	Δ_{Base}	MRR	Δ_{Base}	F1	Δ_{Base}	BA	Δ_{Base}
Qwen3.8-27B													
Base	–	58.99	–	76.04	–	85.71	–	68.69	–	64.84	–	51.62	–
Self-Consistency	ICLR'23	58.53	-0.46	68.20	-7.83	77.42	-8.29	64.98	-3.72	65.86	+1.02	51.98	+0.36
Self-Refine	NeurIPS'23	64.98	+5.99	88.02	+11.98	94.93	+9.22	76.52	+7.83	69.16	+4.32	57.15	+5.53
ExpeL	AAAI'24	79.26	+20.28	85.25	+9.22	92.17	+6.45	83.44	+14.75	80.70	+15.87	70.16	+18.54
MSDiagnosis	KBS'25	59.91	+0.92	87.10	+11.06	**95.39**	+9.68	73.91	+5.22	62.63	-2.21	51.64	+0.02
MetaFlowLLM	NeurIPS'25	63.13	+4.15	82.49	+6.45	88.94	+3.23	73.09	+4.39	74.06	+9.22	62.19	+10.57
ExpRAG	ACL'26	68.20	+9.22	85.25	+9.22	91.24	+5.53	77.04	+8.34	74.19	+9.35	60.84	+9.22
ACE	ICLR'26	59.91	+0.92	74.65	-1.38	85.71	+0.00	68.82	+0.12	63.09	-1.75	50.35	-1.27
ReasoningBank	ICLR'26	58.53	-0.46	78.80	+2.76	86.64	+0.92	69.32	+0.62	66.81	+1.97	53.58	+1.96
MemEvolve	ICML'26	60.83	+1.84	78.80	+2.76	88.02	+2.30	70.87	+2.17	66.50	+1.67	53.22	+1.60
MedGuideEvolve (ours)	Ours	**82.95**	+23.96	**90.32**	+14.29	**95.39**	+9.68	**87.20**	+18.50	**82.17**	+17.34	**73.60**	+21.98
Qwen3.6-27B													
Base	–	64.52	–	76.04	–	82.49	–	71.07	–	67.37	–	54.29	–
Self-Consistency	ICLR'23	66.82	+2.30	74.65	-1.38	80.65	-1.84	71.43	+0.36	70.28	+2.91	57.95	+3.66
Self-Refine	NeurIPS'23	70.51	+5.99	89.40	+13.36	93.09	+10.60	79.40	+8.33	71.78	+4.41	60.80	+6.51
ExpeL	AAAI'24	76.04	+11.52	88.94	+12.90	91.71	+9.22	82.39	+11.32	72.15	+4.78	63.95	+9.65
MSDiagnosis	KBS'25	72.81	+8.29	87.10	+11.06	92.63	+10.14	80.22	+9.16	74.68	+7.31	62.91	+8.62
MetaFlowLLM	NeurIPS'25	76.04	+11.52	88.94	+12.90	92.17	+9.68	82.69	+11.62	76.43	+9.06	68.34	+14.05
ExpRAG	ACL'26	74.19	+9.68	84.33	+8.29	89.86	+7.37	79.85	+8.78	75.80	+8.43	63.27	+8.98
ACE	ICLR'26	67.28	+2.76	77.42	+1.38	81.11	-1.38	72.86	+1.80	67.93	+0.56	55.94	+1.65
ReasoningBank	ICLR'26	67.74	+3.23	80.18	+4.15	84.79	+2.30	74.18	+3.11	71.15	+3.78	59.00	+4.71
MemEvolve	ICML'26	66.82	+2.30	81.57	+5.53	85.25	+2.76	74.05	+2.98	70.06	+2.69	57.84	+3.55
MedGuideEvolve (ours)	Ours	**84.33**	+19.82	**91.71**	+15.67	**94.93**	+12.44	**88.22**	+17.15	**80.29**	+12.92	**76.60**	+22.30
GPT-5.6 Luna Max													
Base	–	58.99	–	75.12	–	87.10	–	68.88	–	63.07	–	52.03	–
ExpeL	AAAI'24	71.89	+12.90	86.18	+11.06	91.24	+4.15	79.31	+10.42	73.08	+10.01	60.93	+8.90
MetaFlowLLM	NeurIPS'25	70.51	+11.52	85.25	+10.14	92.63	+5.53	78.54	+9.65	75.54	+12.47	63.65	+11.62
ExpRAG	ACL'26	80.18	+21.20	87.10	+11.98	91.24	+4.15	84.06	+15.18	79.69	+16.62	68.20	+16.18
ACE	ICLR'26	58.06	-0.92	76.04	+0.92	83.87	-3.23	67.10	-1.78	61.97	-1.10	51.08	-0.95
ReasoningBank	ICLR'26	60.37	+1.38	75.58	+0.46	87.56	+0.46	69.89	+1.01	63.68	+0.61	51.18	-0.85
MemEvolve	ICML'26	58.53	-0.46	75.58	+0.46	85.71	-1.38	67.84	-1.04	65.31	+2.24	52.95	+0.92
MedGuideEvolve (ours)	Ours	**81.57**	+22.58	**89.40**	+14.29	**93.55**	+6.45	**85.87**	+16.99	**80.98**	+17.91	**71.71**	+19.68

contains 940 patient-level cases from 2015 to 2026, split chronologically into 595 Train, 128 Dev, and 217 Test cases; no patient or family member appears in multiple partitions, making the Test split a forward-transfer evaluation. We evaluate a label-blinded diagnostic task using four ranking metrics: Top-1 accuracy, Recall@3, Recall@5, and MRR@5, with Top-1 accuracy serving as the primary endpoint. Top-1 accuracy measures the proportion of cases in which the ground-truth disease is ranked first; Recall@k measures the proportion of cases in which it appears among the top-k candidates; and MRR@5 is the mean reciprocal rank truncated at five, assigning zero when the ground-truth disease is absent from the top five. As secondary class-aware summaries, we additionally report Macro-F1, defined as the unweighted mean of per-class F1 scores, and balanced accuracy, defined as the unweighted mean of per-class recall. All metrics are reported as percentages, and dataset statistics are provided in Appendix D.

Baselines. We evaluate methods across three inference backbones: two open-source models, Qwen3.8-27B (Qwen Team, 2026b) and Qwen3.6-27B (Qwen Team, 2026a), and one proprietary model, GPT-5.6 Luna Max (Singh et al., 2025). We use three baseline groups to isolate gains from current-case reasoning, case/trajectory reuse, and persistent memory: (1) current-case reasoning methods, including Self-Consistency (Wang et al., 2022), Self-Refine (Madaan et al., 2023), and MSDiagnosis (Hou et al., 2025); (2) case- and trajectory-based experience reuse methods, including ExpRAG (Ou et al., 2026), ExpeL (Zhao et al., 2024), and MetaFlowLLM (Fan et al., 2026); and

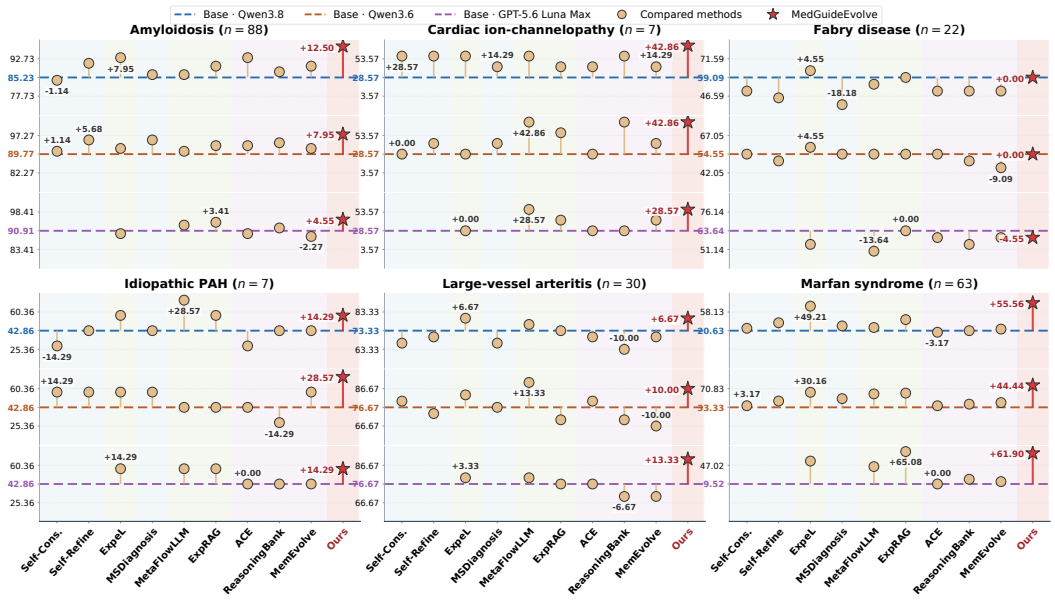


Figure 3: Per-disease Top-1 accuracy on the patient-disjoint Temporal Test across three backbones; bands denote method families, dashed lines show Base, and red stars mark MedGuideEvolve.

(3) persistent memory-based experience reuse methods, including ACE (Zhang et al., 2026), ReasoningBank (Ouyang et al., 2026), and MemEvolve (Zhang et al., 2025). Further implementation details of these baseline methods can be found in Appendix E.

Implementation Details. We serve Qwen3.8-27B and Qwen3.6-27B with vLLM (Kwon et al., 2023) in bfloat16 precision. We set the temperature to 1.0, top-p to 0.95, top-k to 20, min-p to 0, the repetition penalty to 1.0, and the maximum generation budget to 16,384 tokens. Qwen3.8-27B uses the xhigh reasoning setting, whereas Qwen3.6-27B uses the auto reasoning setting. GPT-5.6 Luna Max is accessed through an API with maximum reasoning effort. For experience-based methods, artifacts are constructed offline from the 595 Train cases, method-specific settings are fixed on the 128 Dev cases, and all artifacts are frozen before evaluation on the 217 Test cases; no memory is updated during Test inference. All Qwen runs use two NVIDIA A800 GPUs.

4.2 Main Results

Overall performance. Table 1 establishes a consistent Top-1 advantage for MedGuideEvolve on the patient-disjoint Temporal Test, with accuracy reaching 81.57–84.33% across the three backbones. Relative to Base, gains are 23.96 percentage points on Qwen3.8-27B, 19.82 on Qwen3.6-27B, and 22.58 with GPT-5.6 Luna Max. Experience-reuse comparisons sharpen this result: ExpeL is the closest Top-1 baseline on Qwen3.8, while ExpeL and MetaFlowLLM tie on Qwen3.6; MedGuideEvolve leads by 3.69 and 8.29 points, respectively. With GPT-5.6 Luna Max, it also reaches 81.57% Top-1 and exceeds ExpRAG by 1.38 points. The advantage spans all six ranking and class-aware metrics: MedGuideEvolve is best or tied-best on both Qwen backbones and improves on Base across all six metrics with GPT-5.6 Luna Max. This pattern is consistent with the method's design: validated experience cards connect prior cases to explicit evidence conditions, focusing review on findings relevant to the current patient's differential and organizing that review under established guidelines. Experience thus serves as a context-sensitive aid to evidence review, while guideline-based adjudication remains central. Overall, the results show that these gains extend from first-choice ranking to broader ranking and class-aware performance across backbone settings.

Disease-stratified performance. Figure 3 shows that MedGuideEvolve improves Top-1 accuracy in five of the six disease groups for each backbone. Marfan syndrome records the largest improvement in each setting: accuracy reaches 76.19% on Qwen3.8-27B, 77.78% on Qwen3.6-27B, and 71.43% with GPT-5.6 Luna Max, corresponding to gains of 55.56, 44.44, and 61.90 percentage points, respectively. Amyloidosis offers a complementary profile, improving from a stronger Base

level while maintaining high accuracy across all three backbone settings. Large-vessel arteritis improves across all three backbones, and cardiac ion-channelopathy and idiopathic PAH show the same positive shift; each group includes seven Test cases. Fabry performance matches Base on both Qwen backbones and remains close with GPT-5.6 Luna Max. Together, the figure attributes the aggregate gains to multiple disease groups: Marfan contributes the largest lift, with additional improvements in amyloidosis, arteritis, and the two smaller cohorts across the three backbones. This disease-level profile provides a clinically stratified complement to the aggregate results in Table 1.

4.3 Analysis

The full condition brings coverage and ranking together. Table 2 compares inference-stage ablations with fixed patients and scoring. Guideline-only leads candidate recall among single components, while experience-only leads first-choice accuracy and ranking quality. The full condition produces the strongest ranking profile on both backbones and strong class-aware performance. Together, guidelines broaden coverage while experience sharpens ranking across both backbones.

Table 2: Experience/guideline ablation.

Backbone	Setting	Top-1	Recall@3	MRR@5	Macro-F1
Qwen3.8-27B	Base	58.99	76.04	68.69	64.84
	w/Guideline	69.59	89.40	79.79	74.41
	w/Experience	78.80	85.71	83.56	79.07
	w/full	82.95	90.32	87.20	82.17
Qwen3.6-27B	Base	64.52	76.04	71.07	67.37
	w/Guideline	77.88	90.32	84.67	74.52
	w/Experience	81.57	89.40	85.49	83.77
	w/full	84.33	91.71	88.22	80.29

MedGuideEvolve produces broad patient-level repair gains. Figure 4 shows how each method changes the Base decisions on the same Test patients. MedGuideEvolve shifts substantially more patient decisions toward the correct outcome than away from it on both backbones, yielding the strongest net repair among baselines that reuse cases or trajectories. The net gain exceeds forty cases on each backbone. This patient-level view makes the aggregate gain concrete: the improvement is distributed across individual diagnostic decisions. Because the cohort is shared, the paired design makes the headline gain concrete by showing how individual diagnostic decisions move across both backbones.

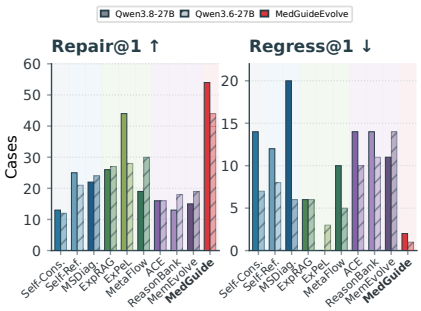


Figure 4: Top-1 transitions: repair and regress.

The improvement persists as the candidate list expands. Figure 5 traces how candidate coverage changes as the list expands. MedGuideEvolve stays at or near the top throughout the displayed range and reaches the highest or tied-highest level at the longest list. The curves rise smoothly for both backbones, showing that the method improves coverage beyond the first choice. Together with the transition plot above, this indicates a consistent benefit: more patients receive a correct first choice while strong alternatives remain available for review. The two panels use independent vertical scales for readability.

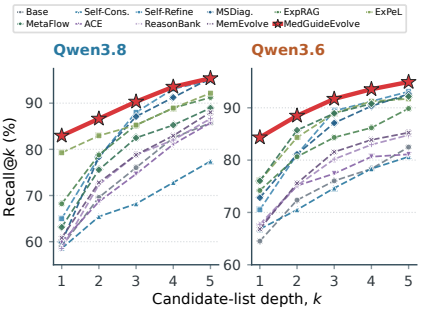


Figure 5: Recall@k by backbone.

5 Conclusion

We introduced MedGuideEvolve, which distills cross-case contrasts into provenance-linked, evidence-conditioned experience cards for diagnostic review. Contrasts among erroneous cases, similar correct cases, and hard controls specify clinical triggers and review actions; independent validation gates each library update. During diagnosis, evidence-matched routing selects the most specific applicable strategy for the patient's evidence and organizes review under established diagnostic guidelines. On the patient-disjoint Temporal Test, MedGuideEvolve improved Top-1 accuracy over Base across all three evaluated backbones. Together, these findings position experience cards as a practical interface between patient evidence and established diagnostic knowledge.

AI USE STATEMENT

Large language models were used as writing assistance tools to help revise the wording and grammar of the authors' draft and improve the clarity of expression.

REFERENCES

Emily Alsentzer, Michelle M. Li, Shilpa N. Kobren, Ayush Noori, Undiagnosed Diseases Network, Isaac S. Kohane, and Marinka Zitnik. Few-shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, 8(1):380, 2025. URL https://www.nature.com/articles/s41746-025-01749-1.

Elena Arbelo, Alexandros Protonotarios, Juan R. Gimeno, Eloisa Arbustini, Roberto Barriales-Villa, Cristina Basso, Connie R. Bezzina, Elena Biagini, Nico A. Blom, Rudolf A. De Boer, et al. 2023 ESC guidelines for the management of cardiomyopathies: Developed by the task force on the management of cardiomyopathies of the european society of cardiology (ESC). *European Heart Journal*, 44(37):3503–3626, 2023. URL https://academic.oup.com/eurheartj/article/44/37/3503/7246608.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. RareAgents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 101–109, 2026. doi: 10.1609/aaai.v40i1.36969. URL https://ojs.aaai.org/index.php/AAAI/article/view/36969.

Shengda Fan, Xin Cong, Zhong Zhang, Yuepeng Fu, Yesai Wu, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Generalizing experience for language agents with hierarchical metaflows. *Advances in Neural Information Processing Systems*, 38:64103–64132, 2026. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5c882988ce5fac487974ee4f415b96a9-Paper-Conference.pdf.

Toyofumi Fujiwara, Yasunori Yamamoto, Jin-Dong Kim, Orion Buske, and Toshihisa Takagi. PubCaseFinder: A case-report-based, phenotype-driven differential-diagnosis system for rare diseases. *The American Journal of Human Genetics*, 103(3):389–399, 2018. doi: 10.1016/j.ajhg.2018.08.003. URL https://doi.org/10.1016/j.ajhg.2018.08.003.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Guangya Yu, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. MSDiagnosis: A benchmark and framework for evaluating large language models in multi-step clinical diagnosis. *Knowledge-Based Systems*, 330:114524, 2025. doi: 10.1016/j.knosys.2025.114524. URL https://doi.org/10.1016/j.knosys.2025.114524.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, and Y. Liu. Agent Hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024. URL https://arxiv.org/abs/2405.02957.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Xiaohao Mao, Yu Huang, Ye Jin, Lun Wang, Xuanzhong Chen, Honghong Liu, Xinglin Yang, Haopeng Xu, Xiaodong Luan, Ying Xiao, et al. A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using EHRs. *npj Digital Medicine*, 8(1):68, 2025. doi: 10.1038/s41746-025-01452-1. URL https://www.nature.com/articles/s41746-025-01452-1.

Justice Ou, Tinglin Huang, Yilun Zhao, Ziyang Yu, Peiqing Lu, Yifei Shen, and Rex Ying. Experience retrieval-augmentation with electronic health records enables accurate discharge QA. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23406–23430, 2026. URL https://aclanthology.org/2026.acl-long.1073/.

Siru Ouyang, Jun Yan, I. Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, et al. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, volume 2026, pp. 94327–94354, 2026.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-27b.

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

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1245. URL https://aclanthology.org/2024.emnlp-main.1245/.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. URL https://arxiv.org/abs/2601.03267.

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. URL https://arxiv.org/abs/2606.09365.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain-of-thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. URL https://arxiv.org/abs/2203.11171.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024. URL https://arxiv.org/abs/2409.07429.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Danyang Zhang, Lu Chen, Situo Zhang, Hongshen Xu, Zihan Zhao, and Kai Yu. Large language models are semi-parametric reinforcement learning agents. *Advances in Neural Information Processing Systems*, 36:78227–78239, 2023. doi: 10.52202/075280-3419. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b22ac37beb5da61efd4882082c9ecd-Abstract-Conference.html.

Guibin Zhang, Haotian Ren, Chong Zhan, Zhenhong Zhou, Junhao Wang, He Zhu, Wangchunshu Zhou, and Shuicheng Yan. MemEvolve: Meta-evolution of agent memory systems. *arXiv preprint arXiv:2512.18746*, 2025. URL https://arxiv.org/abs/2512.18746.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, volume 2026, pp. 86069–86100, 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. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(8106):775–784, 2026. URL https://www.nature.com/articles/s41586-025-10097-9.

A LIMITATIONS

Our method has several limitations. First, whether an experience card enters the library depends on the guideline representation and the cases used for validation. Incomplete guideline coverage or limited validation cases may therefore change the resulting library. Second, the experience library is built offline and remains fixed during Test inference. The current experiments do not evaluate adaptation to revised guidelines or changing clinical workflows. Finally, retrieved cards add prompt context, and building the library incurs additional offline cost. We have not yet quantified the full construction cost or the resulting latency–quality trade-off.

B REPRESENTATIVE GUIDELINE: MARFAN SYNDROME

This appendix presents one selected guideline used by the benchmark: the Marfan syndrome diagnostic guideline. The table below displays the content of the selected guideline in source order, including its clinical content, diagnostic criteria, differential diagnoses, management scope, and diagnostic workflow. Figure 6 shows the diagnostic flowchart provided at the end of the source chapter. This appendix presents the guideline and its diagnostic structure; it does not define a new model, clinical recommendation, or evaluation protocol.

Table 3: Content of the selected Marfan syndrome diagnostic guideline.

Guideline section	Guideline content
Overview	**Clinical content.** Marfan syndrome (MFS) is an autosomal-dominant connective-tissue disorder. The principal systems involved are the skeletal, ocular, and cardiovascular systems. **Diagnostic evidence.** The disorder is commonly associated with pathogenic variation in *FBN1*; a smaller number of cases involve related transforming-growth-factor receptor genes. **Implication.** The diagnostic assessment must integrate findings across multiple organ systems.
Etiology and epidemiology	**Clinical content.** Most affected individuals have a family history, while approximately 15–30% are attributed to de novo variation. Reported prevalence is approximately 1 in 3,000–5,000 individuals. **Diagnostic evidence.** MFS is autosomal dominant. Pathogenic variation in *FBN1* accounts for most molecularly characterized cases; *TGFBR1/2* variation can produce overlapping phenotypes. **Implication.** Family history is an explicit branch in the diagnostic workflow, but its absence does not exclude MFS.
Clinical manifestations: cardiovascular	**Clinical content.** Potential findings include aortic-root dilatation, aortic regurgitation, aortic dissection, and mitral-valve prolapse with or without regurgitation. Symptoms may include dyspnea, chest pain, palpitations, dizziness, syncope, or hemoptysis. **Diagnostic evidence.** Echocardiography and cross-sectional cardiovascular imaging assess the aortic root, ascending aorta, valves, ventricular size, aneurysm, and dissection. **Implication.** Cardiovascular assessment is central because aortic disease drives major morbidity and mortality.

Continued on next page

Table 3 – continued from previous page

Guideline section	Guideline content
Clinical manifestations: skeletal, ocular, and other	**Clinical content.** Skeletal findings may include tall and slender habitus, arachnodactyly, increased arm-span-to-height ratio, scoliosis or kyphosis, pectus deformity, joint laxity, and pes planus. Ocular findings may include ectopia lentis, retinal detachment, cataract, corneal flattening, or axial elongation. Other findings include dural ectasia, striae, recurrent hernia, and a high-arched palate. **Diagnostic evidence.** Ophthalmologic examination, including slit-lamp assessment, evaluates ectopia lentis. The systemic score aggregates characteristic skeletal, skin, ocular, pulmonary, and cardiac findings. **Implication.** No single extra-aortic manifestation is sufficient in isolation; findings are interpreted within the Ghent diagnostic framework.
Auxiliary examinations	**Clinical content.** Recommended assessments include ECG; chest and skeletal radiography; echocardiography; ophthalmologic examination; CT, MRI, or cardiovascular angiography when indicated; and molecular testing. **Diagnostic evidence.** Relevant findings include aortic-root or ascending-aortic dilatation, aortic regurgitation, mitral-valve prolapse, ectopia lentis, and pathogenic variation in *FBN1* or *TGFBR1/2*. **Implication.** The examination set supplies the evidence used by the diagnostic branches below.
Diagnosis: framework	**Clinical content.** The guideline uses the revised 2010 Ghent criteria, combining family history, physical findings, cardiovascular imaging, ophthalmologic examination, and genetic testing. **Diagnostic evidence.** The diagnostic logic differs according to whether a family history of MFS is present. Related disorders, including Loeys–Dietz, Shprintzen–Goldberg, and vascular Ehlers–Danlos syndromes, must be considered and excluded where required. **Implication.** The final diagnosis is a guideline-constrained adjudication, not a similarity-based memory decision.
Diagnosis: no family history	**Clinical content.** MFS may be diagnosed when any one of the following applies: (1) aortic-root Z score ≥ 2 or aortic-root dissection together with ectopia lentis and exclusion of related disorders; (2) aortic-root Z score ≥ 2 or dissection together with a pathogenic *FBN1* variant; (3) aortic-root Z score ≥ 2 or dissection together with systemic score ≥ 7 and exclusion of related disorders; or (4) ectopia lentis together with an aortic aneurysm and a pathogenic *FBN1* variant. **Diagnostic evidence.** The required evidence is conjunctive within each criterion and must be evaluated without treating missing or unperformed examinations as negative findings. **Implication.** If none of the criteria is satisfied, the flow proceeds to exclusion or unresolved assessment rather than automatic diagnosis.
Diagnosis: family history present	**Clinical content.** When a family history of MFS is present, MFS may be diagnosed with ectopia lentis; with systemic score ≥ 7 and exclusion of related disorders; or with an aortic-root Z score ≥ 2 in adults, ≥ 3 in individuals younger than 20 years, or aortic-root dissection, together with exclusion of related disorders. **Diagnostic evidence.** Age-dependent aortic-root thresholds and the family-history branch are preserved explicitly. **Implication.** The family-history branch changes the admissible diagnostic conditions and is not interchangeable with the no-family-history branch.
Systemic score	**Clinical content.** A systemic score of ≥ 7 is diagnostically informative. Scored findings include thumb and wrist signs, pectus deformity, hindfoot deformity or pes planus, pneumothorax, dural ectasia, protrusio acetabuli, reduced upper-to-lower segment ratio or increased arm-span-to-height ratio, scoliosis or thoracolumbar kyphosis, reduced elbow extension, characteristic facial features, striae, severe myopia, and mitral-valve prolapse. **Diagnostic evidence.** The score is an aggregate criterion and must be interpreted together with aortic-root findings, family history, ectopia lentis, genetic evidence, and exclusion of related disorders. **Implication.** The score supports a branch of the diagnostic flow; it does not independently establish MFS.
Differential diagnosis	**Clinical content.** Relevant alternatives include familial thoracic aortic aneurysm and dissection syndromes, congenital contractural arachnodactyly, mitral-valve prolapse syndrome, isolated ectopia lentis, Loeys–Dietz syndrome, Weill–Marchesani syndrome, Shprintzen–Goldberg syndrome, Ehlers–Danlos syndrome, Stickler syndrome, bicuspid aortic-valve disease with aortopathy, homocystinuria, and MASS phenotype. **Diagnostic evidence.** The differential diagnosis is especially important when aortic disease, ocular findings, or systemic features are present without a complete Ghent criterion. **Implication.** An experience card may prompt review of differentiating evidence but cannot remove these exclusion requirements.
Management scope	**Clinical content.** The guideline emphasizes serial aortic surveillance, avoidance of high-intensity or isometric activity, beta-blocker therapy when appropriate, angiotensin-II receptor blockade when tolerated, and surgery for selected aortic or valvular disease. Annual ophthalmologic assessment and management of skeletal or pulmonary complications are also described. **Diagnostic evidence.** The chapter recommends echocardiographic assessment at diagnosis and approximately six months later, followed by surveillance determined by aortic diameter and growth rate. The cited surgical threshold is an aortic diameter of ≥ 50 mm, with earlier intervention considered for rapid enlargement, family history of dissection below 50 mm, or progressive aortic regurgitation. **Implication.** These management recommendations are included for guideline completeness; the present benchmark evaluates diagnosis and candidate ranking rather than treatment decisions.

Continued on next page

Table 3 – continued from previous page

Guideline section	Guideline content
Diagnostic workflow	**Clinical content.** The official workflow begins with clinical suspicion, followed by echocardiography, radiography, orthopedic examination, and ophthalmologic examination. It then branches on family history and evaluates ectopia lentis, aortic-root measurements or dissection, systemic score, pathogenic *FBN1* variation, and exclusion of related disorders. **Diagnostic evidence.** The flowchart below shows the diagnostic workflow provided in the source chapter. **Implication.** The flowchart is the normative reference for the representative guideline shown in this appendix.

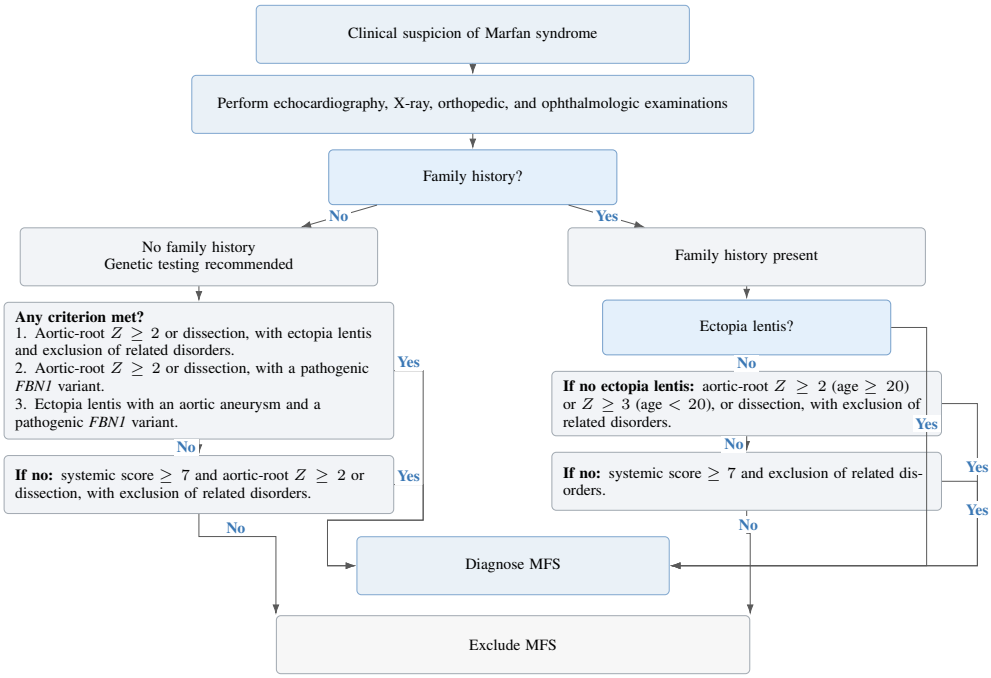


Figure 6: Diagnostic workflow from the selected Marfan syndrome guideline.

C Algorithms and Pseudocode

This appendix specifies the execution order of MedGuideEvolve. Algorithm 1 describes offline experience-library evolution, while Algorithm 2 describes read-only deployment on a label-blinded patient. All symbols follow Section 3. The guideline graph $\mathcal{G}$, model parameters $\boldsymbol{\theta}$, and generation contract Γ remain fixed during both algorithms. Validation outputs and sidecars are locked before reference outcomes and applicability references are revealed; a candidate bundle is committed only after all admission checks pass. The Test-time execution in Algorithm 2 never mutates $\mathbb{M}^{\star}$. The contract-level calls used below denote the corresponding executor operations and do not introduce additional trainable components. Algorithm 1 makes the bundle-level commit boundary explicit, whereas Algorithm 2 exposes the at-most-three-active-path routing budget and round-robin promotion rule.

Algorithm 1 Snapshot–Induce–Admit: Offline Experience-Library Evolution

Input: Locked Case Bank $\mathbb{C}$, initial library $\mathbb{M}_0$, guideline graph $\mathcal{G}$, model $\boldsymbol{\theta}$, generation contract Γ, and fresh patient-disjoint development probe batches $\{\mathbb{P}_t\}_{t=0}^{T-1}$
Output: Frozen admitted library $\mathbb{M}^\star$

$\mathbb{M} \leftarrow \mathbb{M}_0$
for $t = 0, \ldots, T-1$ **do**
  Localize: $\lambda_t \leftarrow \mathrm{Localize}(\mathbb{C}, \mathbb{M})$
  Contrast: $\mathbb{K}_t \leftarrow \mathrm{Contrast}(\lambda_t, \mathbb{C}) = (\mathbb{K}_t^-, \mathbb{K}_t^+, \mathbb{K}_t^{\mathrm{hard}})$ ▷ failure, similar-correct, and hard-control cases
  Induce: $\widetilde{\mathbb{B}}_t \leftarrow \phi(\lambda_t, \mathbb{K}_t, \mathbb{M}; \boldsymbol{\theta}, \Gamma)$
  Project: $\mathbb{B}_t \leftarrow \Pi_{\mathcal{G}}(\widetilde{\mathbb{B}}_t)$ ▷ project, normalize, and attach provenance
  if any schema, provenance, privacy, or guideline check fails **then**
    Reject: record rejection of $\mathbb{B}_t$ and **continue**
  end if
  Snapshot: $\boldsymbol{x}_i \leftarrow \mathrm{Snapshot}(p_i)$ using visible pre-anchor evidence only, for all $p_i \in \mathbb{P}_t$
  Set arms: $\mathbb{U}_{\mathrm{A}} \leftarrow \varnothing$, $\mathbb{U}_{\mathrm{B}} \leftarrow \mathbb{M}$, $\mathbb{U}_{\mathrm{C}} \leftarrow \mathrm{apply}(\mathbb{M}, \mathbb{B}_t)$
  for all $h \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ **do**
    Evaluate: run $\left(\widehat{\boldsymbol{y}}_i^{\,h}, \boldsymbol{s}_i^{\,h}, \boldsymbol{z}_i^{\,h}\right) \leftarrow \mathcal{A}(\boldsymbol{x}_i; \boldsymbol{\theta}, \Gamma, \mathcal{G}, \mathbb{U}_h)$ **for all** $i \in \mathbb{P}_t$
  end for
  Lock: lock all predictions, traces, and sidecars before revealing references
  Score: reveal probe outcomes and independent applicability references
  compute $\mathrm{Net}_3(\mathbb{B}_t)$ and $\mathrm{Net}_1(\mathbb{B}_t)$ from the paired C–B comparison
  record C–A cumulative changes for analysis
  compute card support and unknown rates: $|\mathbb{P}_t^{\mathrm{inj}}(e)| \geq 3$, $|\mathbb{D}_t(e)| > 0$, and $U_t(e)/|\mathbb{D}_t(e)| \leq 0.20$
  Gate: $\mathsf{Pass}_t \leftarrow \mathsf{Schema} \wedge \mathsf{Provenance} \wedge \mathsf{Privacy} \wedge \mathsf{Guideline} \wedge \mathsf{Applicability} \wedge \mathsf{Audit} \wedge \mathsf{Support} \wedge \mathsf{Quantitative}$
    and $\mathrm{Net}_3(\mathbb{B}_t) \geq 1$ and $\mathrm{Net}_1(\mathbb{B}_t) \geq 0$
  if Pass_t **then**
    $\mathbb{M} \leftarrow \mathbb{U}_{\mathrm{C}}$; record the admitted bundle and its provenance
  else
    $\mathbb{M} \leftarrow \mathbb{M}$; record the failed admission and preserve the current library
  end if
end for
$\mathbb{M}^\star \leftarrow \mathbb{M}$
return $\mathbb{M}^\star$

Algorithm 1 makes the admission boundary explicit: C–B provides incremental evidence for the candidate bundle, whereas C–A is retained for the cumulative comparison. An unknown predicate is preserved in the audit trace but does not satisfy an eligibility condition. Thus, a failed check changes neither the current library nor the provenance of earlier admitted bundles.

Algorithm 2 Evidence-Matched Routing and Guideline-Constrained Diagnosis

Input: Label-blinded patient input x, frozen library $\mathbb{M}^\star$, guideline graph $\mathcal{G}$, model $\boldsymbol{\theta}$, generation contract Γ, and node budget $J_{\max}$
Output: Candidate ranking $\widehat{\boldsymbol{y}}$, adjudicated states, and audit trace $\boldsymbol{z}$

Snapshot: $\boldsymbol{x} \leftarrow \mathrm{Snapshot}(x)$ using visible pre-anchor evidence only
$\mathbb{W}^\star \leftarrow \{\mathrm{ver}(e) : e \in \mathbb{M}^\star\}$
Route: $(\widehat{\boldsymbol{y}}^{\,0}, \mathcal{Q}_{\mathrm{active}}, \mathcal{Q}_{\mathrm{wait}}) \leftarrow \mathrm{InitializePaths}(\boldsymbol{x}; \boldsymbol{\theta}, \Gamma, \mathcal{G}), \quad |\mathcal{Q}_{\mathrm{active}}| \le 3$
    Activate up to three supported paths; queue all remaining supported paths.
$\boldsymbol{z} \leftarrow \varnothing$ and $j \leftarrow 0$
while $j < J_{\max}$ and $\mathcal{Q}_{\mathrm{active}} \neq \varnothing$ **do**
    Schedule: $\pi \leftarrow \mathrm{NextActive}(\mathcal{Q}_{\mathrm{active}})$ in round-robin order
    $(\boldsymbol{s}_j, n_j) \leftarrow \mathrm{CurrentState}(\pi)$
    Pool: $\mathbb{R}^\star(\boldsymbol{s}_j) \leftarrow \mathbb{M}^{\star,\mathrm{node}}(\boldsymbol{s}_j) \cup \mathbb{M}^{\star,\mathrm{guide}}(g(\boldsymbol{s}_j)) \cup \mathbb{M}^{\star,\mathrm{global}}$
    Filter: $\mathbb{I}^\star(\boldsymbol{x}, \boldsymbol{s}_j) \leftarrow \{e \in \mathbb{R}^\star(\boldsymbol{s}_j) : \mathrm{use}_{\mathbb{W}^\star}(e \mid \boldsymbol{x}, \boldsymbol{s}_j) = 1\}$
    if $\mathbb{I}^\star(\boldsymbol{x}, \boldsymbol{s}_j) = \varnothing$ **then**
        $e^\star \leftarrow \perp$ ▷ proceed under the established guideline
    else
        Select: $e^\star \leftarrow \arg\max_{e \in \mathbb{I}^\star(\boldsymbol{x}, \boldsymbol{s}_j)}(\mathrm{scope}, \mathrm{overlap}, \mathrm{support}, -\mathrm{order})$
    end if
    Propose: evidence references, rule checks, status, and successor using $(\boldsymbol{x}, \boldsymbol{s}_j, n_j, e^\star; \boldsymbol{\theta}, \Gamma)$
    Adjudicate: $(\boldsymbol{s}_{\mathrm{next}}, n_{\mathrm{next}}, \delta\boldsymbol{z}) \leftarrow \mathcal{T}_{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{s}_j, n_j, e^\star; \boldsymbol{\theta}, \Gamma)$
    Record $(\boldsymbol{s}_{\mathrm{next}}, n_{\mathrm{next}})$ for π in the adjudicated states
    $\boldsymbol{z} \leftarrow \boldsymbol{z} \,\|\, \delta\boldsymbol{z}$; $j \leftarrow j + 1$
    if $j \ge J_{\max}$ **then**
        Stop all: set $\mathcal{Q}_{\mathrm{active}}, \mathcal{Q}_{\mathrm{wait}} \leftarrow \varnothing$
    else if $\mathrm{TerminalOrBlocked}(\boldsymbol{s}_{\mathrm{next}}, n_{\mathrm{next}})$ or $\mathrm{RepeatedFingerprint}(\boldsymbol{s}_{\mathrm{next}}, n_{\mathrm{next}})$ **then**
        Stop: remove π from the active queue and promote the next supported path from the wait queue, if available
    else
        Continue: update π to $(\boldsymbol{s}_{\mathrm{next}}, n_{\mathrm{next}})$ and requeue it
    end if
end while
Finalize: $\widehat{\boldsymbol{y}} \leftarrow \mathrm{Finalize}(\widehat{\boldsymbol{y}}^{\,0}, \text{adjudicated states}; \Gamma)$
return $(\widehat{\boldsymbol{y}}, \text{adjudicated states}, \boldsymbol{z})$

Algorithm 2 is read-only with respect to $\mathbb{M}^\star$ and $\mathcal{G}$. Similarity may be used to form the routing pool, but it never grants eligibility: scope, version, and evidence predicates determine whether a card is injected. The model may propose evidence references and a transition, but $\mathcal{T}_{\mathcal{G}}$ recomputes the evidence state from the immutable snapshot and applies the guideline guards before appending the transition to the audit trace.

D Dataset Details

CardioRareBench is a de-identified dataset of 940 patient-level cases collected at a single hospital. The patient-level case is the unit of analysis. The six disease categories evaluated in the chronological Dev and Temporal Test partitions are amyloidosis, large-vessel arteritis, cardiac ion-channelopathy, Fabry disease, idiopathic pulmonary arterial hypertension, and Marfan syndrome. Two additional categories, coronary artery ectasia and glycogen storage disease, occur only in Train and are retained for offline experience formation. Table 4 summarizes the dataset scope.

Table 4: CardioRareBench overview. Train-only categories occur only in the Train partition.

Item	Description
Dataset	CardioRareBench
Unit of analysis	De-identified patient-level case
Total cases	940
Evaluation categories	Amyloidosis; large-vessel arteritis; cardiac ion-channelopathy; Fabry disease; idiopathic pulmonary arterial hypertension; Marfan syndrome
Train-only categories	Coronary artery ectasia; glycogen storage disease

Table 5: Chronological partitioning and experimental use of CardioRareBench.

Partition	Cases	Time window	Experimental role	Label/result usage
Train	595	2015–2023	Offline experience formation	Used for experience construction
Dev	128	2024	Method selection and candidate validation	Used to select and fix settings
Test	217	2025–2026	Final temporal evaluation	Used for scoring only after prediction is complete

Table 6: Representative case from the Temporal Test partition. The reference disease is not shown to the model.

Field	Case content
Demographics	63-year-old man.
Chief complaint	Chest tightness for 2 years; nocturnal orthopnea for 3 months; symptoms worsened during the preceding week.
Clinical course	Progressive dyspnea, lower-extremity edema, and a 5-kg weight gain over the preceding month. Previous evaluation reported positive immunofixation electrophoresis and bone-marrow aspiration showing a small clonal plasma-cell proliferation. Outside testing reported cardiac troponin T 0.09, BNP 277, decreased T3/T4, creatinine 72, and platelets 66. Chest CT showed bilateral infection, cardiomegaly, partial left-upper-lobe atelectasis, and lower-esophageal varices. Coronary CT angiography showed mild LAD/LCX stenosis. Echocardiography showed right-heart and left-atrial enlargement, left-ventricular wall thickening, diffuse left-ventricular hypokinesis, left-ventricular dysfunction (EF 46%), elevated pulmonary artery systolic pressure, and right pleural effusion. The patient was unable to lie flat because of dyspnea and edema.
Past medical history	Hepatitis B cirrhosis for 20 years; splenic artery embolization more than 10 years earlier; impaired coagulation. No known hypertension or diabetes.
Electrocardiography	Sinus rhythm, complete right bundle-branch block, right-axis deviation, and T-wave changes; QRS duration 168–180 ms and QT interval 456–482 ms.
Laboratory findings	Total protein 59; free-light-chain ratio 0.074; lambda free light chain 451; kappa free light chain 33.5; high-sensitivity C-reactive protein 16.3; LDH 584 decreasing to 508; high-sensitivity cardiac troponin T 0.473 increasing to 0.656; NT-proBNP 6397 increasing to 8612. Units and reference ranges follow the source record and are omitted here.
Reference disease	Amyloidosis.

Cases are partitioned chronologically by index year, with no patient or family member appearing in multiple partitions. Table 5 summarizes the temporal windows and roles: Train forms offline experience, Dev fixes method settings, and Temporal Test is used only for final scoring.

The model receives each instance as a complete, de-identified, label-blinded case record containing the available pre-anchor clinical evidence. Empty fields and internal identifiers are omitted from the display below for readability, while the clinical content is preserved in substance. Table 6 presents one representative Temporal Test case; its reference disease is shown only as a paper annotation and is not included in the model input.

E Baselines and Metrics

E.1 Implementation of Baseline Methods

The implementation details of the baseline methods are as follows. All baselines receive the same de-identified, label-blinded patient case and produce an ordered five-slot disease candidate list under the generation settings described in Section 4.1.

Self-Consistency. Self-Consistency samples multiple reasoning paths and selects the most frequent final answer by majority voting. In our diagnostic-task adaptation, we sample five independent responses for each case under the same generation contract and select the diagnosis receiving the most Top-1 votes. Ties are resolved using the total occurrence count of each candidate across all five slots, mean confidence, replicate order, slot order, and normalized candidate name. The aggregated response is converted back to the common five-slot output schema before evaluation.

Self-Refine. Self-Refine uses a three-step answer–feedback–revision procedure. The model first produces an initial disease ranking, then provides feedback on the ranking, and finally revises the ranking using the original case and the feedback. Only the final revised JSON response is used for evaluation.

MSDiagnosis. We adapt the MSDiagnosis procedure to the diagnostic task using three steps. The model first produces a forward candidate ranking, then performs backward verification for each candidate against the relevant diagnostic criteria, and finally refines the ranking according to the verification results. The verification step distinguishes necessary, supportive, exclusion, and confirmation criteria. Evidence marked as not tested or not recorded is not treated as negative evidence.

ExpRAG. For each case, we flatten the structured clinical record into a sequence of clinical evidence statements and use the resulting representation to retrieve similar cases from the Train case bank. We construct a BM25 index over the Train cases and retrieve the top five cases for each input. Each retrieved case is represented by a truncated clinical case card together with its confirmed diagnosis. The retrieved case cards are appended to the model context before the final disease ranking is generated.

ExpeL. We adapt ExpeL by extracting reusable diagnostic insights from successful and unsuccessful Train trajectories. The abstraction procedure forms success–failure pairs and maintains an insight set through four operations: ADD, EDIT, UPVOTE, and DOWNVOTE. Newly extracted insights receive an initial importance score; supporting and contradicting trajectories increase or decrease this score, respectively. Insights whose importance reaches zero are removed. The final insight list is rendered as a frozen diagnostic experience card and appended to the model context for subsequent cases.

MetaFlowLLM. We adapt MetaFlowLLM by converting Train trajectories into hierarchical diagnostic workflows. Each workflow represents a reusable sequence of diagnostic review steps, including intermediate evidence checks and the information required by subsequent steps. The workflows are constructed offline from the Train trajectories and merged into a frozen hierarchical workflow card. At inference time, the workflow card is appended to the patient case context, and the model produces the final disease ranking under the common output schema.

ACE. We construct an ACE-style playbook from Train trajectories by distilling reusable diagnostic actions and decision rules. The resulting entries are stored in a structured memory artifact. For a new patient case, we retrieve the most relevant entries using BM25 similarity and append the retrieved playbook entries to the model context before diagnosis.

ReasoningBank. ReasoningBank is implemented as a repository of reusable diagnostic strategies extracted from previous Train trajectories. Each strategy records a compact reasoning pattern that can be applied to similar clinical cases. At inference time, the current case is used to retrieve the most relevant strategies from the frozen repository, and the retrieved strategies are included as additional context for disease ranking.

MemEvolve. MemEvolve starts from an initial set of diagnostic memory entries and applies memory mutation operations to add, edit, or remove reusable strategies. Candidate memory architectures are selected using the Dev split, after which the selected memory is frozen for Test evaluation. During inference, the current case retrieves the most relevant memory entries, which are appended to the model context before the final disease ranking is generated.

E.2 METRICS

We evaluate the ordered disease candidate lists on the 217 Temporal Test cases. Top-1 accuracy measures whether the reference disease is ranked first. Recall@k measures whether the reference disease appears among the first k candidates, and MRR@5 assigns reciprocal-rank credit when the reference disease appears within the first five positions. Macro-F1 and balanced accuracy are computed from the top-ranked disease over the six disease categories represented in the Temporal Test partition.

Candidate names are normalized against the fixed disease ontology before scoring. Malformed or missing candidate outputs are treated as invalid predictions rather than being silently removed from the evaluation denominator. All metrics are reported as percentages, and changes from Base are reported in percentage points.

F PROMPTS

Final Inference Prompt

```
[SYSTEM]

You are a clinician responsible for generating diagnostic candidates for the retrospective
      diagnosis of a cardiovascular rare-disease case.

Answer solely based on the clinical_record in the user message. Unrecorded information
     does not imply a negative finding; do not fabricate test results, medical history, or
      treatments.

Based on the evidence actually provided in the case, output exactly five disease-level
     diagnostic candidates in descending order of likelihood. Candidate names must be
     diseases, syndromes, or disease subtypes. Do not use symptoms, signs, examination
     items or results, complications, treatments, treatment effects, disease stages, or "
     differential diagnosis" itself as candidate names.

For each candidate, provide an integer confidence score from 0 to 10 and 1 to 3 concise
     sentences of diagnostic reasoning grounded in the evidence from the case. Rank the
     candidates from highest to lowest confidence. Do not read or infer hidden labels.

Output exactly one valid JSON object. Do not output Markdown, code fences, or any
     additional text. Do not output case_id. The output format must be exactly:
{"most_likely_diagnosis":{"diagnosis1":{"diagnosis_name":"string","confidence_score":0,"
     diagnostic_reasoning":"string"},"diagnosis2":{"diagnosis_name":"string","
     confidence_score":0,"diagnostic_reasoning":"string"},"diagnosis3":{"diagnosis_name":"
     string","confidence_score":0,"diagnostic_reasoning":"string"},"diagnosis4":{"
     diagnosis_name":"string","confidence_score":0,"diagnostic_reasoning":"string"},"
     diagnosis5":{"diagnosis_name":"string","confidence_score":0,"diagnostic_reasoning":"
     string"}}}

[USER]

{
  "clinical_record": {clinical_record}
}
```

Candidate Experience Skill Construction Prompt

```
[USER]
[GUIDELINE-CONSTRAINED INDUCTION CONTEXT]
You construct a candidate bundle of conditional Experience Skills for
MedGuideEvolve. The bundle will be projected against the established
diagnostic guideline and evaluated on fresh patient-disjoint probes before it
can enter the frozen Experience Library. Do not make an admission decision in
this prompt.

[LOCALIZED PROBLEM]
Guideline location or node:
{guideline_location}
Error type:
{error_type}
Visible evidence context:
{visible_evidence_context}
[CURRENT EXPERIENCE LIBRARY]
{current_experience_library}
Use the current library only to refine or complement existing skills. Do not
copy a skill unless the new contrast supports its scope and predicates.

[ORDERED CONTRAST]
[FAILURE CASES: K^-]
{failure_cases}
[SIMILAR CORRECT CASES: K^+]
{similar_correct_cases}
[HARD CONTROLS: K^hard]
{hard_control_cases}
The contrast contains recurring failure cases, similar correctly handled cases,
and hard controls that share salient evidence but support a different
diagnostic direction or applicability boundary. Use all three groups:

1. Failure cases identify the recurring guideline-application error.
2. Similar correct cases identify the evidence and review action associated
   with the desired decision.
```

```
3. Hard controls define when the proposed skill must not be applied.

Construct a finite candidate bundle of reusable Experience Skills. Each skill
must connect a guideline location and role to explicit applicability predicates
and a scoped review action. The skill must guide evidence review; it must not
replace the established guideline or directly invent a diagnosis.

For every candidate skill, specify:

- location: the guideline or diagnostic node where the skill applies;
- role: the guideline-application role, such as evidence review, differential
  comparison, or confirmation review;
- scope: exactly one of node, guideline, or global, using the most specific
  scope justified by the contrast;
- required_predicates: evidence conditions that must evaluate to true;
- exclusion_predicates: evidence conditions that must evaluate to false;
- review_action: the conditional action the executor should perform;
- provenance: the non-identifying source references, guideline version, and
  problem context supporting the skill.

Predicates must preserve the following assertion states when applicable:
present, absent, not_done, not_recorded, and conflicting. Do not treat
not_recorded as absent. Every predicate must be linked to supporting evidence
in the supplied trajectories or guideline context. Hard controls should be
used to define exclusion predicates or applicability boundaries rather than
as positive evidence.

Before returning the bundle, make it projection-ready:

- use only guideline-supported locations, roles, evidence concepts, and
  confirmation or exclusion requirements;
- remove unsupported or contradictory predicates and actions;
- ensure required and exclusion predicates do not conflict;
- keep each skill conditional, patient-transferable, and specific enough for
  evidence-matched routing;
- reject a candidate when its evidence is insufficient, its scope is too broad,
  or its action would bypass a required guideline guard.

Return exactly one JSON object with the following schema:

{
  "candidate_bundle": [
    {
      "location": "guideline_or_node",
      "role": "evidence_review_or_differential_or_confirmation",
      "scope": "node|guideline|global",
      "required_predicates": [
        {
          "concept": "string",
          "assertion_state": "present|absent|not_done|not_recorded|conflicting",
          "evidence_links": ["reference"]
        }
      ],
      "exclusion_predicates": [
        {
          "concept": "string",
          "assertion_state": "present|absent|not_done|not_recorded|conflicting",
          "evidence_links": ["reference"]
        }
      ],
      "review_action": "string",
      "provenance": {
        "problem_context": "string",
        "source_references": ["reference"],
        "guideline_version": "string",
        "guideline_location": "string"
      }
    }
  ]
}

Use an empty candidate_bundle when the contrast does not support a
guideline-consistent transferable skill. Do not output Markdown, case IDs,
patient-identifying information, hidden labels, a final diagnosis ranking, an
admission decision, or any text outside the JSON object.
```