EvoTD: Decoupled Memory Evolution for Specialty Triage and Diagnostic Inference

Anonymous authors
Paper under double-blind review

Abstract

Clinical practice requires clinicians to synthesize heterogeneous specialty-specific knowledge before making diagnostic decisions. Medical question answering has therefore become a widely employed task for such knowledge synthesis, owing to its controllable evaluation. However, most prior approaches either treat specialty triage as a one-time preprocessing step without leveraging it during diagnostic inference or update notes without enforcing consistency with the selected specialty-specific context. To this end, we propose EvoTD, a decoupled memory evolution framework for specialty triage and diagnostic inference. Specifically, DualEvo is first designed to maintain separate specialty-boundary patches and notes, preventing diagnostic-experience updates from altering specialty triage. Then, SkillRoute adaptively selects deterministic boundary rules, graph-based scoring, and LLM reasoning to construct specialty-specific context via case evidence. To improve the reliability of memory evolution, we develop ReliHarness to synthesize entity-indexed notes from predictions and harness their usage through warmup, single-pass pre-injection, and harmful-memory quarantine. Comprehensive experiments show that EvoTD improves over direct prompting by 7.1 and 6.1 points (3-seed means) on NEJMQA and MMLU-Pro Health, respectively.

1 Introduction

Specialty-aware medical artificial intelligence (AI) systems must synthesize heterogeneous clinical evidence and specialty-specific knowledge before producing a diagnostic inference. This requirement matters most when a question can be approached from several specialties, each prioritizing different evidence. Medical question answering (QA) provides a controlled setting for testing whether a system applies clinically relevant knowledge rather than merely recognizing a disease label. Large language model (LLM)-based systems perform strongly on this task (Singhal et al., 2023; Tu et al., 2024) and fall into three paradigms (Figure 1a–c). The first encodes medical knowledge in fixed parameters (Singhal et al., 2023; Tu et al., 2024): effective when test cases match the optimization regime, degraded by shifts in task format or case distribution. The second augments a general LLM with external static knowledge at inference time (Cui et al., 2025; Jeong et al., 2024; Asai et al., 2024); because that source is fixed for a given deployment, the context is insufficient whenever relevant evidence is absent or not retrieved. The third accumulates experience from interaction trajectories without updating parameters (Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2023a), letting earlier feedback inform later reasoning, but applied to multi-stage clinical workflows, it leaves the ownership and stage-specific reuse of experience unspecified.

Across these paradigms, we identify an unresolved coupling between specialty triage and diagnostic inference. Triage selects the clinical perspective and the specialty-specific context for the downstream inference, yet existing workflows treat that assignment as weak preprocessing, so inference may underuse or override the selected context, and notes are updated without verifying consistency with the context that conditioned the inference. Both behaviors yield decisions that are internally inconsistent across the two stages.

To address these challenges, we propose EvoTD, a decoupled memory evolution framework for specialty triage and diagnostic inference. DualEvo maintains separate specialty-boundary patches and diagnostic notes under a one-way triage-to-diagnosis constraint, so that diagnostic experience never directly re-labels the specialty perspective. SkillRoute then builds specialty-conditioned

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

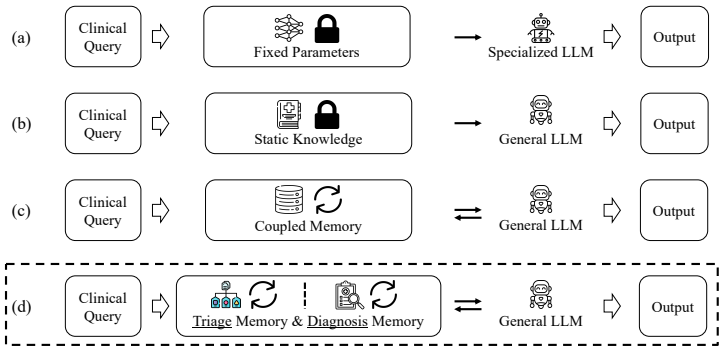

Figure 1: Framework comparison. Lock: knowledge not updatable at inference; circular arrow: updatable memory. (a) Specialized LLM from fixed parameters. (b) General LLM plus static task knowledge. (c) General LLM reading and updating one coupled memory. (d) EvoTD keeps triage and diagnosis memory separate, evolved under a one-way triage-to-diagnosis constraint.

context at an evidence-adaptive depth, resolving clear cases with deterministic boundary rules and reserving skill-injected LLM reasoning for ambiguous ones. Finally, ReliHarness synthesizes entity-indexed notes from graded traces and governs their use through warmup, single-pass pre-injection, and harmful-memory quarantine, turning feedback into traceable and auditable updates rather than unconstrained memory edits. On NEJMQA and MMLU-Pro Health with DeepSeek-v4-Flash as the backbone, EvoTD improves diagnostic accuracy over direct prompting by $+7.1$ and $+6.1$ points (3-seed means), respectively.

In summary, our contributions are fourfold:

- We formulate specialty triage and diagnostic inference as a coupled decision process with decoupled evolution channels, exposing cross-stage memory interference as a source of inconsistent inferences.
- We propose DualEvo to separate specialty-boundary patches from diagnostic note updates, so that patches condition downstream inference while notes never enter the router.
- We develop SkillRoute to construct specialty-conditioned context at an evidence-adaptive depth using deterministic boundary rules, graph-based scoring, and skill-injected LLM reasoning.
- We introduce ReliHarness to make memory evolution auditable and regression-aware through graded-trace harvesting, warmup, single-pass pre-injection, and harmful-memory quarantine.

2 Related Work

Specialty Triage and Clinical Agents. Domain-specialized large language models show that medical knowledge can be encoded in model parameters (Singhal et al., 2023; Tu et al., 2024), while multi-agent and workflow-based systems organize complementary clinical roles (Kim et al., 2024; Wang et al., 2025; Bao et al., 2025; Peng et al., 2025) and dedicated benchmarks evaluate them, from simulated clinical environments to medical multimodal perception (Schmidgall et al., 2024; Liu et al., 2026a; Ye et al., 2026; Peng et al., 2026). Specialty triage instead identifies the perspective from which a case should be assessed, through explicit and reproducible protocols (Gilboy et al., 2020), LLM-based natural-language triage (Masanneck et al., 2024; Levine et al., 2024), or clinical knowledge graphs linking case evidence to structured knowledge (Cui et al., 2025). These establish strong components for clinical reasoning, but none address how a specialty assignment stays operative during later diagnostic inference as the system evolves from feedback. That setting exposes an asymmetry — a misroute replaces the departmental context wholesale, whereas a diagnostic memory misfires only on the cases it matches — so the two stages warrant different acceptance rules.

Experience-Driven Evolution of LLM Agents. Building on reasoning–acting agents (Yao et al., 2023), recent systems convert feedback and interaction traces into reusable guidance: Self-Refine revises outputs from self-feedback (Madaan et al., 2023), Reflexion stores verbal feedback from

failed trajectories (Shinn et al., 2023), ExpeL extracts and retrieves generalizable insights (Zhao et al., 2024), AutoGuide generates state-conditioned guidelines from offline experience (Fu et al., 2024), Voyager builds a skill library (Wang et al., 2023a), SkillWeaver and Agent Workflow Memory induce reusable skills and workflows (Zheng et al., 2025; Wang et al., 2024), FireAct distills trajectories into parameters (Chen et al., 2023), and R-Zero studies co-evolution (Huang et al., 2026). A parallel line manages the memory itself, retrieving over a memory stream (Park et al., 2023) or consolidating long-term stores (Packer et al., 2024; Zhong et al., 2024). In medical settings, EvoClinician applies test-time evolutionary learning (He et al., 2026), EvoMDT evolves multi-agent decision processes (Liu et al., 2026b), and GSEM and SkillOpt retrieve or edit stored experience (Han et al., 2026; Yang et al., 2026); beyond text, MedExpMem carries experience reuse into multimodal settings such as MRI enhancement, brain-tumor world models, and functional MRI dynamics (Lin et al., 2026; Wang et al., 2026; Yuan et al., 2026; Feng et al., 2026). These demonstrate the value of non-parametric adaptation, and distilling traces into conditionally retrieved guidance is by now an established move that ExpeL and AutoGuide already occupy; we do not claim that move as our contribution. What this line does not model is the distinct *ownership* and downstream effect of triage-boundary patches versus diagnostic memories, nor the resulting asymmetry in what an update is permitted to change. DUALEVO isolates the two channels so that patches may alter the context supplied to diagnostic inference, whereas diagnostic memories cannot alter specialty triage, and RELIHARNESS assigns each channel its own acceptance rule. Stage ownership together with asymmetric gating, not trace distillation or retrieval, is what we ask to be judged as novel.

Adaptive Context Use and Reliable Memory Evolution. Retrieval-augmented methods supply external evidence without modifying parameters (Xu et al., 2024), selecting strategies by question complexity (Jeong et al., 2024), jointly learning retrieval and self-critique (Asai et al., 2024), grounding decisions in knowledge graphs (Cui et al., 2025; Sun et al., 2024; Luo et al., 2024), or allocating extra inference only when needed (Fan et al., 2025). These select inference-time evidence rather than controlling how feedback-derived knowledge accumulates. That control is what self-evolving systems lack. Apparent self-improvement can regress after further updates (Wu et al., 2025), a generation–verification gap can limit it (Song et al., 2025), and agents can deviate through memory and workflow changes (Shao et al., 2026). Three consequences shape our design. First, when teacher and predictor are the same model, asking it to repair a case it has just failed is unreliable (Huang et al., 2024), so the learning signal should come from traces where the knowledge was actually applied and verification should draw on evidence outside that model (Gou et al., 2024). Second, re-eliciting an answer is itself stochastic — the property chain-of-thought (CoT) prompting and self-consistency exploit (Wei et al., 2022; Wang et al., 2023b) — so a corrective second pass can overwrite answers that were already right. Third, a memory that raises mean accuracy can still overturn correct answers, so an update must be accepted or rejected on the individual regressions it causes rather than on the average it moves. SKILLROUTE and RELIHARNESS address these requirements, the former building specialty-conditioned context at an evidence-adaptive depth and the latter governing both channels through warmup, single-pass pre-injection, replay verification, and harmful-memory quarantine.

3 PROPOSED METHOD

3.1 OVERALL FRAMEWORK

Figure 2 shows the framework. A medical QA case is first converted into a specialty-conditioned context by SKILLROUTE, then answered by the diagnostic model, and finally written into a TRIAGECARD trace store from which TriageEvo and DiagnosisEvo update separate memories. The stages differ in responsibility: triage decides *which* clinical perspective the case is read under, diagnosis reasons *within* it.

3.2 PROBLEM DEFINITION

For an item $x_t = (q_t, \mathcal{O}_t)$, let Θ_R and Θ_D denote the frozen routing and diagnostic prompts, Δ_t the patch state, $\mathcal{N}_t$ the note pool, and $\mathcal{K}$ the clinical knowledge graph. The route–diagnose path is

$$d_t, \mathcal{C}_t = \mathcal{R}(x_t; \Theta_R, \Delta_t, \mathcal{K}), \tag{1}$$

$$\hat{y}_t = \mathcal{D}(x_t, \mathcal{C}_t, \mathcal{G}(\mathcal{N}_t; x_t, d_t, t); \Theta_D)\,. \tag{2}$$

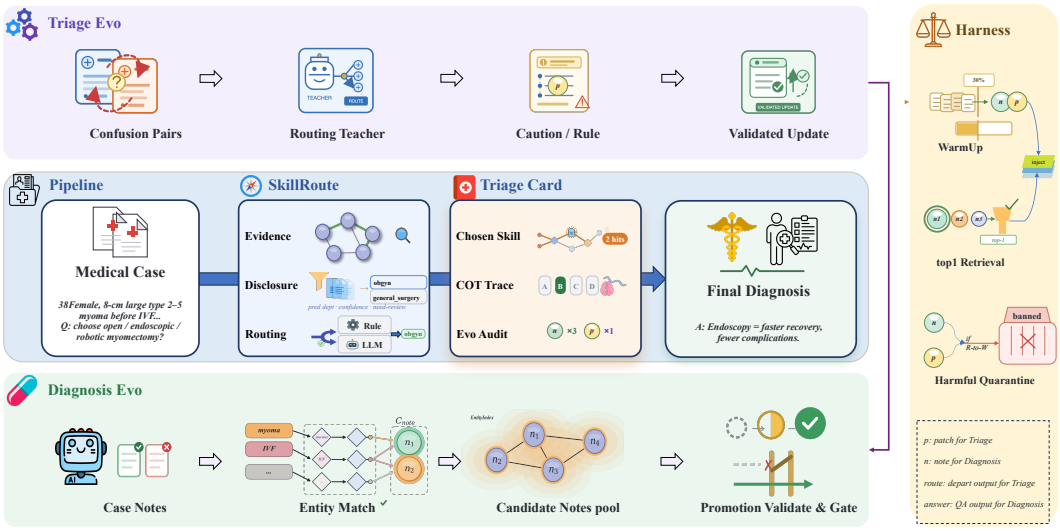

Figure 2: EvoTD architecture. DualEvo evolves asymmetrically: TriageEvo repairs label boundaries under replay gates; DiagnosisEvo builds an entity-keyed note pool with harmful-memory quarantine. Both share a TriageCard trace store; notes never enter the router.

Here d_t is the selected specialty, $\mathcal{C}_t$ is the specialty-conditioned context, and $\mathcal{G}$ is the reliability harness gate for note injection. After prediction, the system records an auditable trace

$$\tau_t = \textsc{TriageCard}(x_t, e_t, d_t, \mathcal{C}_t, \mathcal{M}_t, \hat{y}_t, r_t, s_t), \tag{3}$$

where e_t is the extracted evidence frame, $\mathcal{M}_t$ the injected note identifiers, r_t the diagnostic rationale, and s_t the execution spans of evidence extraction, skill retrieval, routing, verification, and prediction. The gold label is appended only after prediction, so the trace supports scoring and memory evolution without leaking answer labels into the forward pass.

3.3 Decoupled Dual Self-Evolution

DualEvo treats triage and diagnosis as an inference-coupled but evolution-decoupled process, following the acceptance asymmetry set out above. Both channels read the same TriageCard history, but only TriageEvo can change the specialty boundary used by future routing. Notes may condition the diagnostic model, but they are excluded from routing. Every other input of Eq. 1 is held fixed, so the constraint removes $\mathcal{N}_t$ alone and leaves the knowledge graph $\mathcal{K}$ on both sides:

$$\mathcal{R}(x; \Theta_R, \Delta_t, \mathcal{K}, \mathcal{N}_t) = \mathcal{R}(x; \Theta_R, \Delta_t, \mathcal{K}). \tag{4}$$

This one-way constraint is the central reliability boundary: it keeps the medically useful route-to-diagnosis dependency while barring notes from re-labeling the specialty perspective.

TriageEvo. TriageEvo monitors recurrent specialty-confusion pairs in the trace store and proposes triage patches when enough cases expose the same ambiguity. A candidate patch δ is accepted only when replay over a reference set $\mathcal{B}_t^R$ shows zero observed regression. $\mathcal{B}_t^R$ holds the most recent 10 LLM-path traces that were both routing- and answer-correct, since a caution text can change an LLM decision but not a deterministic one:

$$\text{Accept}_T(\delta) = \mathbb{I}\left[\sum_{\bar{\tau} \in \mathcal{B}_t^R} \mathbf{1}\{\mathcal{R}_{\Delta_t \oplus \delta}(x_{\bar{\tau}}) \neq \mathcal{R}_{\Delta_t}(x_{\bar{\tau}})\} = 0\right]. \tag{5}$$

The strict gate reflects the asymmetric cost of routing errors, because a bad specialty boundary can recondition many future diagnostic contexts, whereas a bad note is matched only to a narrower set of cases. One residual coupling remains: $\mathcal{B}_t^R$ is filtered on answer correctness, so notes can change which traces enter it; Eq. 4 constrains the routing function and its inputs, not the patch gate's composition.

DiagnosisEvo. DiagnosisEvo distills graded diagnostic traces into entity-indexed advisory memories, each storing a source specialty, an answer entity, an applicability pattern, a distilled reasoning rule,

source-trace identifiers, and a harm flag. Unlike TriageEvo, this channel is intentionally local, since a note applies only when the target case matches its specialty and entity pattern, and it is retired once its incorrect injections outnumber its correct ones.

Why the one-way boundary matters. Notes are indexed by $(\textit{dept}, \textit{gold_entity})$ from graded traces. Were they visible to the router, a later item mentioning the same entity could recover its specialty tag from note metadata rather than from clinical reasoning. Equation 4 removes that path by construction.

What the entity index does and does not read. The gold-entity key comes from *training* traces, where the label is legitimately observed; at test time none exists and none is consulted. Retrieval scores a note against the item's visible stem and option strings alone, and the specialty filtering the pool is the router's own prediction d_t, not the gold specialty. The label reaches the TRIAGECARD only after the prediction is emitted, so it shapes memory for *later* items but never the current forward pass; a note fires when its entity appears among the *candidate* options, decidable without knowing which is correct.

3.4 ADAPTIVE SKILL-INJECTION DEPTH

SKILLROUTE constructs the specialty-conditioned context before note injection. It uses the same route skill at different depths rather than separate routers, where deterministic boundary rules handle clear cases, graph and cue evidence handle medium-confidence cases, and skill-injected LLM reasoning is reserved for ambiguous cases. Depth is therefore monotone in routing uncertainty and in conflict among specialty signals, so that low boundary-rule confidence — an elicited quantity, hence usable but imperfect (Xiong et al., 2024) — or disagreement between graph and cue evidence triggers deeper disclosure of specialty decision cards to the LLM, while agreement keeps routing deterministic or graph-scored; the prompt budget thus tracks ambiguity instead of being fixed.

Given the selected specialty, SKILLROUTE assembles context from four sources: department answer priors, professional facts, knowledge-graph highlights, and accepted patches. Professional facts require sufficient lexical support, while knowledge-graph highlights and boundary cautions are tied to the routed specialty. This gives a meaningful cold start, since before any note exists the model already receives case-specific context from rules, graph evidence, and decision cards; only after warmup does RELIHARNESS add notes as diagnostic hints.

3.5 CLINICAL RELIABILITY EVOLUTION HARNESS

RELIHARNESS controls how feedback becomes non-parametric memory. It uses the TRIAGECARD store as an audit log and applies different risk gates to the two evolution channels: two filters act on DiagnosisEvo before prediction (warmup, then entity-matched top-1 retrieval) and a third is channel-agnostic (harmful-memory quarantine). We write the candidate filter and top-1 retrieval compactly as

$$\mathcal{P}_t = \mathrm{Filter}(\mathcal{N}_t; d_t, x_t, \mathrm{harm}),$$
$$n_t^\star = \mathrm{Top1}_{n\in\mathcal{P}_t}\, \mathrm{sim}(n, x_t),$$

where Filter keeps only memories with the same specialty, a matching answer entity, and no harmful flag. When $\mathcal{P}_t = \emptyset$ the $\mathrm{Top1}$ operator has no argument; we then set $\mathcal{G} = \emptyset$ and the item is answered from SKILLROUTE context alone, so an empty pool degrades to the no-memory configuration rather than to an undefined injection. The active note set is

$$\mathcal{G}(\mathcal{N}_t; x_t, d_t, t) = \begin{cases} \{n_t^\star\}, & t \geq w|\mathcal{S}| \wedge \mathrm{dir}(n_t^\star, x_t), \\ \emptyset, & \text{otherwise}, \end{cases} \tag{6}$$

so warmup, an entity-matched candidate and direction support are all required. Quarantine is channel-specific, because the two memories are credited differently. Write $\mathrm{good}(m)$ and $\mathrm{bad}(m)$ for a note's injections on which the emitted answer was correct and incorrect, and $\mathrm{W2R}(p)$ and $\mathrm{R2W}(p)$ for the wrong-to-right and right-to-wrong events a patch is credited with from the trace store. Then

$$\mathrm{Quar}(m) = \begin{cases} \mathbb{I}[\mathrm{bad}(m) \geq 1 \ \wedge\ \mathrm{bad}(m) > \mathrm{good}(m)], & m \in \mathcal{N}_t, \\ \mathbb{I}[\mathrm{R2W}(m) \geq 2 \ \wedge\ \mathrm{R2W}(m) > \mathrm{W2R}(m)], & m \in \Delta_t. \end{cases} \tag{7}$$

Notes are credited at outcome level, not by counterfactual flips: attributing a flip would require re-predicting without the note, the very call single-pass injection removes. The stricter patch threshold compensates for its wider blast radius.

The harness follows five auditable stages. *Trace*: each forward pass writes the full TRIAGECARD of Eq. 3. *Synthesize*: both channels distill from traces where the model's knowledge is observable rather than asking the same model to invent knowledge it has just failed to apply — in a pilot of 10 error cases with five generations each, eight remained incorrect throughout (95% Clopper–Pearson CI $[44\%, 97\%]$). TriageEvo therefore proposes patches (p) from recurrent specialty confusions and DiagnosisEvo distills entity-level rules from graded traces as notes (n), both governed as *memories* by the shared harness (Fig. 2, right). *Gate*: because a regression can outweigh a missed improvement, warmup suppresses injection over the first 30% of the stream, top-1 entity retrieval keeps only the highest-relevance memory, and the quarantine flag of Eq. 7 retires a memory once its observed harm outweighs its observed help. *Verify*: acceptance is prospective — a patch deploys only after zero-regression replay, and a candidate whose source cases overlap that replay set is rejected, so no patch is confirmed by the traces that proposed it; a note must come from a graded trace whose answer was already correct and match the target's specialty and entity. *Inject*: the accepted memory is inserted once before a single prediction as a clinical reasoning hint, avoiding the extra stochastic call a corrective second pass would add — a choice Table 4 tests directly.

4 EXPERIMENTS

Implementation. DeepSeek-v4-Flash, with thinking disabled and temperature 0, serves as both note-generating teacher and test-time predictor, making self-evolution same-model by default. Two variants probe the backbone: DeepSeek-v4-Pro predicts from Flash-written notes, whereas GLM-5.1 both writes and consumes its own. Four nested configurations isolate each component: Pure LLM, SKILLROUTE (adaptive routing and context, no evolution), +TriageEvo, and +DUALEVO (full pipeline with RELIHARNESS). External baselines run their original pipelines on the same items, and Reflexion is reported at its label-free first trial because its native stopping rule consults the test label before retrying. Every accuracy cell in Table 1 is a mean±sample standard deviation over three seeds; Pure LLM is deterministic at temperature 0 and shown as a single value. Hyper-parameters: warmup ratio w=0.3, triage-patch threshold of 5 accumulated confusions, note pruning every 20 cases, and boundary fast-path confidence c>0.9.

Main results. EVOTD improves diagnostic accuracy over direct prompting on both benchmarks in all four prompt-style × inference-mode settings, and it does not fall below its own backbone in any of them (Table 1). The comparison fixes the predictor to DeepSeek-v4-Flash and pairs every method against the Pure LLM run of *its own* setting, so prompt style and thinking mode are never confounded with the method effect. The contrast is sharpest under thinking mode on NEJMQA, where the Pure LLM is already strong at 79.42 and MedExpMem, MDAgents, Reflexion and SkillOpt all fall *below* it; only GSEM (80.88) and EVOTD (80.4–81.4) clear that bar, and GSEM sits above our structured variant, so we do not claim the strongest absolute accuracy in that cell. The separation we do claim is cost: GSEM needs roughly an order of magnitude more calls per item (Figure 3).

Magnitude and significance. The gains are largest where the backbone is weakest. On NEJMQA no-think EVOTD lifts the structured baseline from 68.0 to 75.14 and the CoT baseline to 79.5, and on MMLU-Pro Health it leads every column. Each three-seed mean in Table 3 is tested with an item-level bootstrap and a paired sign-flip permutation test, avoiding the false precision of concatenating repeated runs over identical items. Both mean gains exclude zero in their bootstrap intervals, and those positive lower bounds matter clinically, since the gain is directional rather than an average of large offsetting changes; wrong-to-right corrections outnumber right-to-wrong regressions in both blocks, so evolution rarely overturns an already-correct diagnosis.

Online end-to-end evolution. Table 2 moves to the harder *online end-to-end* regime: EVOTD starts empty and accumulates every patch and note online from per-case feedback over the full corpus, and it still leads against strong open- and closed-source LLMs. On MMLU-Pro Health it is best on all three reported metrics, ahead of the strongest closed model (Gemini 2.5 Flash) and its own backbone; on NEJMQA it leads on accuracy and MCC while Gemini retains F_1 (Table 2). We do not read the margins over the strongest external rows as a resolved ranking: they are of the same order as the run-to-run spread in Table 1 and we run no paired test against those systems. What the table supports

Method	Source	NEJMQA R-Acc	NEJMQA A-Acc	MMLU-Pro R-Acc	MMLU-Pro A-Acc
No-think					
Pure LLM (struct)	—	83.6	68.0	77.24	74.80
Pure LLM (CoT)	—	83.2	74.2	78.9	75.61
MedExpMem	[MICCAI'26]	–	71.99±2.40	–	77.92±1.69
MDAgents	[NeurIPS'24]	–	68.99±2.06	–	75.34±0.94
Reflexion	[NeurIPS'23]	–	72.40±3.79	–	76.02±0.00
GSEM	[arXiv'26]	–	78.42±1.44	–	80.35±0.85
SkillOpt	[arXiv'26]	–	77.73±2.79	–	79.81±2.08
EvoTD (struct)	Ours	83.33±0.83	75.14±2.26	**80.50±0.00**	**80.89±0.82**
EvoTD (CoT)	Ours	**84.0±0.83**	**79.5±0.85**	80.37±0.23	79.54±0.23
Think					
Pure LLM (struct)	—	83.2	79.42	79.3	76.02
Pure LLM (CoT)	—	84.8	79.42	78.5	76.42
MedExpMem	[MICCAI'26]	–	78.69±0.41	–	79.67±0.70
MDAgents	[NeurIPS'24]	–	75.68±0.95	–	75.75±2.01
Reflexion	[NeurIPS'23]	–	76.50±1.66	–	76.29±0.62
GSEM	[arXiv'26]	–	80.88±1.25	–	81.16±1.02
SkillOpt	[arXiv'26]	–	78.96±2.26	–	81.71±0.82
EvoTD (struct)	Ours	**86.20±0.89**	80.41±1.03	**80.83±0.06**	**81.90±1.67**
EvoTD (CoT)	Ours	85.80±0.89	**81.40±0.92**	79.83±0.29	81.42±1.03

Table 1: Main results on each corpus's test split.

Config	EvoTD	ΔA	W2R/R2W	95% CI	p
NEJMQA, no-think structured (ref. Pure LLM 68.03)					
seed 42	76.23	+8.20	35/15	[+2.46, +13.93]	0.0066
seed 43	72.54	+4.51	23/12	[+0.00, +9.43]	0.0895
seed 44	76.64	+8.61	29/8	[+3.69, +13.52]	0.0008
3-seed mean	**75.14±2.26**	**+7.10**	**87/35**	[+2.87, +11.48]	**0.0016**
MMLU-Pro Health, no-think structured (ref. Pure LLM 74.80)					
seed 42	81.71	+6.91	20/3	[+3.25, +10.57]	0.0005
seed 45	80.08	+5.28	19/6	[+1.63, +9.35]	0.0146
seed 46	80.89	+6.10	19/4	[+2.44, +9.76]	0.0026
3-seed mean	**80.89±0.82**	**+6.10**	**58/13**	[+2.71, +9.76]	**0.0007**

Table 3: Paired significance of EvoTD's diagnostic-accuracy gain against pure LLM reference.

Method	A-Acc	F1	MCC
NEJMQA			
Claude Haiku 4.5	64.58±1.06	64.75±1.02	52.73±1.38
GPT-5.1	76.28±1.47	76.28±1.48	68.37±1.93
Gemini 2.5 Flash	78.57±0.73	**78.69±0.62**	71.36±0.94
Phi4 14B	55.88±10.18	55.91±10.28	42.55±11.23
Deepseek-R1 32B	61.63±1.04	61.69±1.04	48.80±1.38
DeepSeek-v4-Flash	79.61±0.18	77.25±0.10	73.47±0.21
WizardLM 8x22 141B	56.39±2.54	57.04±2.17	42.34±3.11
Debate [ICML'23]	66.41±1.61	64.86±2.66	56.54±1.57
SelfMoA [arXiv'25]	46.16±6.18	49.62±8.40	31.34±9.91
MoA [ICLR'25]	47.33±6.10	51.32±3.07	34.08±4.42
EvoTD	**80.72±0.79**	78.43±1.14	**74.91±1.07**
MMLU-Pro Health			
Claude Haiku 4.5	68.77±1.85	68.74±1.83	65.18±2.06
GPT-5.1	71.35±3.87	71.45±3.87	68.09±4.32
Gemini 2.5 Flash	77.34±0.67	77.34±0.68	74.77±0.76
Phi4 14B	70.46±0.99	70.40±0.98	67.07±1.09
Deepseek-R1 32B	62.71±2.98	63.29±2.26	58.64±3.07
DeepSeek-v4-Flash	78.47±1.81	78.46±1.83	76.00±2.02
WizardLM 8x22 141B	60.68±8.89	61.19±7.94	56.57±9.23
Debate [ICML'23]	67.93±0.86	67.95±0.83	64.33±0.95
SelfMoA [arXiv'25]	60.19±11.26	60.86±10.04	55.99±11.90
MoA [ICLR'25]	65.31±7.26	65.57±6.76	61.38±7.98
EvoTD	**80.97±0.27**	**80.99±0.29**	**78.79±0.30**

Table 2: Online end-to-end evolution, no split: EvoTD makes one streaming pass from an empty memory, answering each case from within-stream memory; other rows are static references. Per block: first three closed-source, rest open-source, per MAC's protocol (Peng et al., 2025).

is that a single streaming pass from empty memory is competitive with strong static predictors, not that it dominates them.

Inference efficiency. EvoTD's accuracy is also the cheapest to obtain at serving time. Under think mode it holds the Pareto frontier of both corpora at about 1.6 LLM calls per item, whereas GSEM reaches comparable accuracy only at ∼13 calls, an 8× cost (Figure 3). The saving comes from single-pass pre-injection, which spends no extra self-critique or retrieval round; the routing call itself can be replaced by a structured decision model at equal accuracy and lower cost (Further study, below). Two qualifications bound the claim: the axis counts serving-time calls only, offline note synthesis being amortized over the development split, and the points are three-seed means without dispersion. The figure therefore supports a lower inference-call count, not an unconditional total-cost advantage.

Controlled ablations. Three single-run ablations on the frozen NEJMQA test split localize where the gain originates and why it must be screened; the flow comparison uses CoT, and a separate stream is marked where applied.

Component attribution. SkillRoute is the *cold-start substrate* and reference floor for every evolution claim: all ΔA values in Table 4 (middle) and Table 5 are measured against the paired SkillRoute run, not direct prompting. The nested block cites direct prompting only to expose the substrate's headroom (+3.28) — context construction, not a claimed contribution.

Within that block the evolution channels carry the claim. Adding TriageEvo moves answer accuracy, inside the routing variation stated above and consistent with its role of stabilising the boundary rather than answering; DiagnosisEvo supplies the remaining +4.92, for +8.61 in total. The accompanying $p = 0.016$ tests the full stack against the SkillRoute floor and so does *not* isolate TriageEvo's increment. A separate offline stream isolates DiagnosisEvo alone at +3.05 over its own paired baseline (18/8, $p = 0.078$), corroborating the direction on independent data; in that stream, toggling the diagnosis channel moves routing accuracy by only −0.61 points, as Eq. 4 predicts for a channel the router cannot read. We therefore claim TriageEvo as a reliability and auditability component, not an accuracy lever: the incremental test that would settle its accuracy contribution — DiagnosisEvo first, then TriageEvo — is not among the runs reported here.

One-way versus bidirectional flow. Table 4 (middle) contrasts the one-way constraint with a symmetric variant that also feeds notes back into the router. The objection to that variant is architectural: a router reading a note keyed on an entity recovers the note's specialty tag instead of reasoning about the case, so any routing gain is not evidence of better triage. The numbers agree but are small and

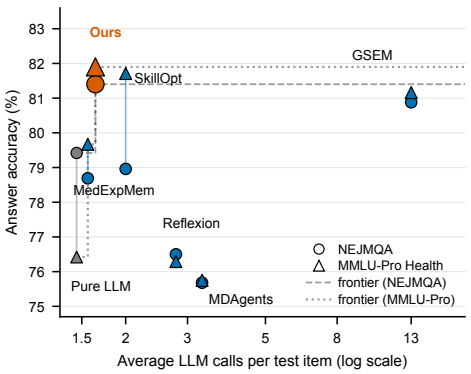

Figure 3: Answer accuracy versus *serving-time* LLM calls per item, think mode, three-seed means. Grey indicates the direct prompting; blue indicates baselines; vermillion: EvoTD (Pareto frontiers, $\sim$1.6 calls/item); dispersion omitted for legibility, and offline note synthesis is amortized over the development split, not counted.

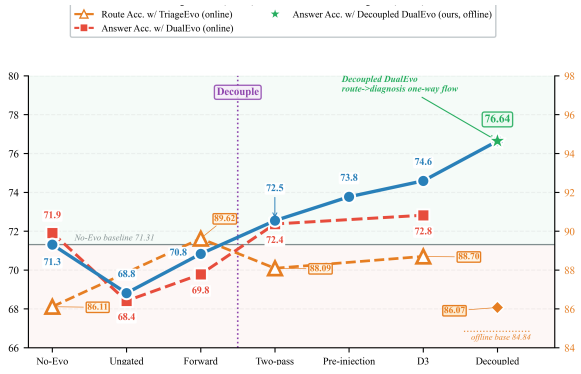

Figure 4: Design history of the diagnosis channel, not a controlled comparison: stages were measured under the protocol current when built, so stage heights are not mutually comparable. Answer accuracy left, routing accuracy right; unscreened stages regress, screening order separates improvement from regression. X-axis order is chronological, not a controlled factor and supports no causal interpretation.

inside the run-to-run variation above (Table 4, middle), so we present the constraint as a reliability property, not an accuracy lever.

Reliability mechanisms. Dropping warmup costs 2.0 points (76.64→74.59) and lowers the flip ratio from 19/6 to 13/5 (Table 4, bottom). Replacing single-pass pre-injection with a two-pass predict-then-correct step collapses accuracy to 68.44, essentially the Pure LLM level (68.03), and inverts the flip ratio to 11/18 because the second pass rewrites more correct answers than it repairs. Figure 4 shows the same lesson as a design trajectory. Screening order, not the mere presence of memory, separates improvement from regression, and unscreened evolution is worse than none. Warmup also shapes the online regime, where the entity-keyed pool fills only after the warmup fraction has passed, so a streaming pass spends its early cases without memory support, a coverage limit we leave to future work.

Where the gain comes from. Two cases illustrate the intended behavior. In an enhanced-recovery question, injecting a note stating that patient-controlled analgesia is the recommended postoperative regimen resolves a scattered prediction to the correct option; in a cardiac-myxoma case, a note indexed to *aortic stenosis*, a distractor option, does not pull the prediction toward it. A note thus acts as a reasoning cue when relevant, while a surface entity match does not by itself steer the answer, consistent with the +3.0 open-ended gain where the match key is the model's own prediction (Table 5).

Robustness and scope. Table 5 varies the backbone. Evolution improves the paired routed baseline in all three settings, but accuracy and margin fall once the predictor is no longer Flash (Table 5), so teacher–predictor transfer remains open.

Why the model is its own note teacher. Because teacher and predictor are the same model, DiagnosisEvo can be read as an effect similar to self-distillation while remaining train-free: supervision comes from the model's own graded traces and is stored in context rather than weights. It also explains why a stronger predictor does not automatically enlarge the gain: residual errors are dominated by content the model can already express but applies inconsistently. We claim only this weaker reading, not parametric distillation or the injection of new medical knowledge.

Open-ended generalization. The second boundary removes answer options altogether. We run DiagnosisEvo on MedQA-Open (Nachane et al., 2024), an option-free, single-specialty reformulation of MedQA (Jin et al., 2021), matching notes through the model's own first-pass predicted entity; TriageEvo does not apply without labels, making this the hardest setting. Even so, accuracy rises from 53.8% to 56.8% across three seeds (+3.0 points on average; Table 5), with descriptive flip counts across the three runs favoring improvement (38 W2R versus 8 R2W). No individual seed reaches statistical significance (exact McNemar $p = 0.54, 0.076$, and 0.21). We do not pool discordant pairs

TriageEvo	DiagnosisEvo	A-Acc	ΔA	W2R/R2W	95% CI	p
✗	✗	71.31	+3.28	—	[−4.92, 11.48]	0.480
✓	✗	71.72	+3.69	—	—	—
✗	✓‡	74.09	+3.05	18/8	[0.00, 6.10]	0.078
✓	✓	**76.64**	**+8.61**	**19/6**	**[1.64, 9.43]**	**0.016**

Info flow	ΔR	ΔA	W2R/R2W	95% CI	p
Triage → Diagnosis	+0.82	+5.33	45/32	[−2.05, 12.70]	0.172
Triage ↔ Diagnosis	+1.23	+4.10	50/40	[−3.69, 11.89]	0.343

Mechanism removed	A-Acc	ΔA	W2R/R2W	95% CI	p
Warmup	74.59	+6.56	13/5	[0.00, 6.56]	0.096
Pre-injection (two-pass)	68.44	+0.41	11/18	[−7.38, 1.23]	0.265

Table 4: Three controlled ablations on NEJMQA (single runs): nested channels; one-way vs. bidirectional flow; one RELIHARNESS mechanism removed. Δ vs. Pure LLM; each CI/p tests that row against its SKILLROUTE predecessor, so none isolates TriageEvo. Middle-block intervals include zero, indicating uncertainty in estimated gains. ‡Separate stream, own baseline.

Backbone and teacher, NEJMQA						
Predictor	**Teacher**	**R-Acc**	**A-Acc**	**ΔA_{pair}**	**R-MCC**	**A-MCC**
Flash	Flash	85.25	76.64	+5.33	.819	.694
Pro	Flash	84.02	72.54	+1.23	.803	.640
GLM-5.1	GLM-5.1	82.44	74.81	+2.29	.785	.670

Open-ended transfer, MedQA-Open				
Seed	**Base**	**+Evo**	**95% CI**	**McNemar p**
44	55.2	57.0	[−2.4, +5.9]	0.54
45	52.2	56.5	[+0.0, +8.7]	0.076
46	53.9	57.0	[−1.0, +7.1]	0.21
Mean	53.8	56.8	—	—

Table 5: Two scope boundaries, single runs. Top: backbone and teacher (GLM-5.1: single-specialty subset). Bottom: open-ended transfer. McNemar tests are conducted separately for each seed. ΔA_{pair} vs. paired SKILLROUTE.

across seeds because the same items are evaluated repeatedly. We therefore interpret the result as a consistent directional improvement rather than a statistically significant effect; matching requires one additional retrieval pass.

Further study. As a routing-side complement, we replaced the only generative routing call with a structured System One decision model (Jev Model) on all 244 NEJMQA test items: answering a five-way choice and a binary unambiguity question in one request, it matches the generative router within two cases (R-Acc 0.873 vs. 0.865) at $3.7\times$ lower cost and $4.5\times$ lower latency. Enriching its decision state with extracted evidence and KG retrieval does not improve the routing choice but repairs the binary verification signal (misroute AUC $0.52 \rightarrow 0.79$); full arms, calibration and paired tests are in the supplementary appendix.

5 DISCUSSION

Why decoupled evolution. The two channels differ in update risk, so patches are rejected after any observed replay regression whereas notes are only quarantined upon observed harm. Keeping the flow one-way closes a workflow-level *misevolution* pathway (Shao et al., 2026) and removes a confound in which specialty-tagged notes would inflate routing accuracy (Table 4).

What the gain does and does not show. That a train-free, same-model pipeline improves diagnostic accuracy *suggests* the residual bottleneck here is partly activation and consistency rather than raw knowledge (Song et al., 2025); it does not show that the mechanism repairs reasoning the model cannot perform.

Limitations and ethics. (1) The knowledge-gap pilot has only ten error cases. (2) Both benchmarks are multiple-choice examinations, so they measure neither open-ended diagnosis, acuity, calibration, outcomes, nor real-world routing. (3) Answer entities recur between development and test at very different rates (37.7% versus 100%), so generalization to unseen entities is untested; because notes are keyed by entity this is the sharpest threat to the diagnostic channel, and we report no entity-disjoint partition, shuffled-key control or plain entity-cache baseline that would separate reasoning reuse from lookup, nor claim to have separated them. (4) Streaming evaluation needs gold labels after prediction, whereas real feedback may be delayed, noisy or absent; replay also detects regressions only on retained cases, giving no guarantee under shift. EVOTD is accordingly not intended for clinical deployment, for the reasons set out in the supplementary ethics statement.

6 CONCLUSION

We presented EVOTD: patches and notes evolve asymmetrically under a one-way triage→diagnosis constraint, with per-case specialty-conditioned context and harm-screened updates. Diagnostic accuracy improves over direct prompting by 3-seed means of $+7.1$ and $+6.1$ points on the NEJMQA and MMLU-Pro Health benchmarks, respectively, mostly from DiagnosisEvo; the one-way constraint buys reliability, and unscreened evolution proves worse than none.

AI USE STATEMENT

Generative AI tools were used in two narrow roles. First, for language editing of this manuscript: tightening phrasing, correcting grammar, and harmonising terminology. Second, for software-engineering assistance on the reference implementation, including code restructuring, debugging, and writing test and evaluation scaffolding.

Every AI-assisted output was read in full, executed, and corrected by the authors before being kept; nothing was accepted on the basis of the tool's assertion alone. All numerical results reported in this paper were produced by logged runs of the authors' own code over the released corpora, and no reported figure, table entry, accuracy, confidence interval, or p-value was generated, estimated, or modified by an AI tool.

AI tools were not used to produce any figure or image in this paper. Every figure was either drawn by hand in presentation software or rendered by the Python/matplotlib scripts included in the supplementary code package, which are the scripts that generated the exact PDFs embedded here. AI tools were also not used to generate or alter data, gold labels, split assignments, experimental results, the research questions, the experimental design, or the interpretation of results. Each bibliographic entry was checked by the authors against the cited source. The authors take full responsibility for the content of this submission.

ETHICS STATEMENT

This work evaluates language-model systems on publicly available multiple-choice medical examination benchmarks. No human subjects, no protected health information and no new data collection were involved. EVOTD is a research prototype and is not intended for clinical deployment: its memory evolves without human review, and RELIHARNESS makes updates auditable and reversible rather than clinically safe. Clinical use would require prospective validation, human oversight and regulatory compliance. We have no competing interests to declare.

REPRODUCIBILITY STATEMENT

Every number in this paper can be regenerated from the supplementary code package, which contains: the full routing, dual-evolution and harness implementation; both evaluation corpora together with their exact train/test partitions and identifiers; the answer prompts and the few-shot exemplars used at inference; end-to-end runners for the offline and the online regime; and the scripts that render each figure. Seeds are fixed in the runners and written into every emitted record, and the runners emit the per-case and per-seed logs from which each reported mean, confidence interval and paired significance test is aggregated, so those quantities are recomputable rather than only quotable; the logs themselves are generated artefacts and are not part of the submitted archive. The backbone, decoding settings and all hyper-parameters (warmup ratio, triage-patch trigger threshold, patch-acceptance replay rule, note-pruning interval, boundary fast-path confidence, top-1 note retrieval) are stated in the main text or the appendix, and no value was tuned on the test partitions. Reproduction requires calls to proprietary LLM endpoints, whose server-side behaviour can drift over time, so a re-execution need not reproduce a logged run exactly; the shipped seeds and prompts bound the variation a reader should expect. On acceptance the code package will be released under an open-source licence.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Liuxin Bao, Zhihao Peng, Xiaofei Zhou, Runmin Cong, Jiyong Zhang, and Yixuan Yuan. Expertise-aware multi-llm recruitment and collaboration for medical decision-making. *arXiv preprint arXiv:2508.13754*, 2025.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. FireAct: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023.

Hejie Cui, Jiaying Lu, Ran Xu, Shiyu Wang, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, Liang Zhao, et al. A review on knowledge graphs for healthcare: Resources, applications, and promises. *Journal of Biomedical Informatics*, 169:104861, 2025. doi: 10.1016/j.jbi.2025.104861.

Wei Fan, JinYi Yoon, and Bo Ji. iMAD: Intelligent multi-agent debate for efficient and accurate LLM inference. *arXiv preprint arXiv:2511.11306*, 2025.

Qianhan Feng, Zhongzhen Huang, Yakun Zhu, Yannian Gu, Winnie Chiu Wing Chu, et al. MedExp-Mem: Adapting experience memory for differential diagnosis. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2026. Early Accept. arXiv:2605.22872.

Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. Autoguide: Automated generation and selection of context-aware guidelines for large language model agents. In *Advances in Neural Information Processing Systems*, 2024.

Nicki Gilboy, Paula Tanabe, Debbie Travers, and Alexander M. Rosenau. Emergency severity index (ESI): A triage tool for emergency department care, version 4. Technical report, Agency for Healthcare Research and Quality, Rockville, MD, 2020.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024.

Xiao Han, Yuzheng Fan, Sendong Zhao, Haochun Wang, and Bing Qin. GSEM: Graph-based self-evolving memory for experience augmented clinical reasoning. *arXiv preprint arXiv:2603.22096*, 2026.

Yufei He, Juncheng Liu, Zhiyuan Hu, et al. EvoClinician: A self-evolving agent for multi-turn medical diagnosis via test-time evolutionary learning. *arXiv preprint arXiv:2601.22964*, 2026.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-Zero: Self-evolving reasoning LLM from zero data. In *International Conference on Learning Representations*, 2026.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 7029–7043, 2024.

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

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. *arXiv preprint arXiv:2404.15155*, 2024.

David M. Levine, Rudraksh Tuwani, Benjamin Kompa, Amita Varma, Samuel G. Finlayson, Ateev Mehrotra, and Andrew Beam. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: An observational study. *The Lancet Digital Health*, 6:e555–e561, 2024. doi: 10.1016/S2589-7500(24)00097-9.

Yiyang Lin, Chenhui Wang, Zhihao Peng, and Yixuan Yuan. Unifield: A unified field-aware MRI enhancement framework. *arXiv preprint arXiv:2603.09223*, 2026.

Peigen Liu, Rui Ding, Yuren Mao, Ziyan Jiang, Yuxiang Ye, Yunjun Gao, Ying Zhang, Renjie Sun, Longbin Lai, and Zhengping Qian. OpenHospital: A thing-in-itself arena for evolving and benchmarking LLM-based collective intelligence. *arXiv preprint arXiv:2603.14771*, 2026a.

Qicai Liu, Zhichao Hu, Tao Huang, Yupeng Niu, Xinche Zhang, Shanwu Ma, Chutong Lin, Kim Huat Goh, Hyeokkoo Eric Kwon, Feng Gao, et al. EvoMDT: A self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *npj Digital Medicine*, 9:124, 2026b. doi: 10.1038/s41746-025-02304-8.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, 2024.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lars Masanneck, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven G. Meuth, Lennert Böhm, and Marc Pawlitzki. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study. *Journal of Medical Internet Research*, 26:e53297, 2024. doi: 10.2196/53297.

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. Few shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 542–573, 2024. doi: 10.18653/v1/2024.findings-emnlp.31.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. In *arXiv preprint arXiv:2310.08560*, 2024.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology*, 2023.

Zhihao Peng, Liuxin Bao, and Yixuan Yuan. Mac: Masked agent collaboration boosts large language model medical decision-making, 2025. URL `https://arxiv.org/abs/2507.21159`.

Zhihao Peng, Cheng Wang, Shengyuan Liu, Zhiying Liang, Zanting Ye, Min Jie Ju, Peter YM Woo, and Yixuan Yuan. Omnibrainbench: A comprehensive multimodal benchmark for brain imaging analysis across multi-stage clinical tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42732–42743, June 2026.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Shuai Shao, Qihan Ren, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song, Linfeng Zhang, Weinan Zhang, Dongrui Liu, and Jing Shao. Your agent may misevolve: Emergent risks in self-evolving LLM agents. In *International Conference on Learning Representations*, 2026.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham M. Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *International Conference on Learning Representations*, 2025.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, 2024.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654*, 2024.

Chenhui Wang, Boyun Zheng, Liuxin Bao, Zhihao Peng, Peter YM Woo, Hongming Shan, and Yixuan Yuan. Brain-wm: Brain glioblastoma world model. *arXiv preprint arXiv:2603.07562*, 2026.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

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. In *International Conference on Learning Representations*, 2023b.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. MedAgent-Pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ting Wu, Xuefeng Li, and Pengfei Liu. Progress or regress? self-improvement reversal in post-training. In *International Conference on Learning Representations*, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. In *International Conference on Learning Representations*, 2024.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Yuqing Yang, Dongdong Chen, Xue Yang, and Chong Luo. SkillOpt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Zanting Ye, Xiaolong Niu, Xuanbin Wu, Xu Han, Shengyuan Liu, Jing Hao, Zhihao Peng, Hao Sun, Jieqin Lv, Fanghu Wang, et al. Unveiling and bridging the functional perception gap in mllms: Atomic visual alignment and hierarchical evaluation via pet-bench. *arXiv preprint arXiv:2601.02737*, 2026.

Yixuan Yuan, Yu Jiang, Cheng Wang, Zhihao Peng, Zhibin He, Boyun Zheng, Yifei Sun, Tuo Zhang, Peter Yat Ming Woo, Robert D Sanders, et al. Modeling individual-level long-range brain dynamics from short fMRI scans. *Research Square preprint*, 2026. doi: 10.21203/rs.3.rs-9217734/v1.

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, 2024.

Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, et al. SkillWeaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024.