FROM CASE FEEDBACK TO AUDITABLE MEMORY: CONTEXT EVOLUTION FOR COMPLETE BLOOD COUNT-BASED RISK-ORIENTED REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) show promise in medical reasoning, yet accumulating reusable and auditable experience from structured laboratory findings across case streams remains challenging. To address these issues, we propose a feedback-driven context evolution framework that adapts frozen LLMs through structured memory and inference time context updates. Specifically, we first construct a hierarchical memory architecture that separates stable knowledge from dynamic experience and organizes audit events, memory cards, and retrieval indices to preserve clinical guidance and traceable case-derived experience. We then design a deterministic write gate that checks consistency, merges compatible rules, isolates conflicts, and routes ambiguous candidates for human review, thereby controlling how new experience enters memory. At inference time, section-first retrieval dynamically adjusts the number and composition of retrieved cards to the configured context capacity, integrating relevant stable knowledge with case-derived experience. We evaluate the framework on complete blood count-based case streams for disease detection, tumor risk assessment, and candidate tumor group ranking. Compared with retrieval-augmented generation and other context optimization baselines, the framework improves disease detection and candidate tumor group ranking, while exhibiting sustained performance across a long-horizon case stream.

1 INTRODUCTION

Large language models (LLMs) have shown strong potential in clinical applications such as medical question answering, report interpretation, and decision support (Singhal et al., 2023; 2025; Thirunavukarasu et al., 2023; Yang et al., 2022). Among these settings, reasoning over structured laboratory findings presents a particularly constrained form of clinical reasoning: models must interpret correlated numerical abnormalities, distinguish clinically similar patterns, and produce conclusions that remain grounded in the limited evidence available (McCoy et al., 2025; Bhasuran et al., 2025). Unlike open-ended medical question answering, laboratory-based reasoning therefore requires the model to preserve fine-grained evidence boundaries rather than relying on fluent but weakly supported associations. These characteristics make structured laboratory case streams a useful setting for studying how reasoning experience can be accumulated and reused without compromising previously reliable knowledge.

Existing approaches improve medical LLMs through continued pretraining, supervised fine-tuning, parameter-efficient adaptation, retrieval-augmented prompting, or external memory (Lewis et al., 2020; Wei et al., 2022; Hu et al., 2022; Shinn et al., 2023; Zhang et al., 2026). Weight-updating methods can improve domain competence but are costly to refresh and difficult to audit when knowledge or data conditions change. Retrieval- and memory-augmented systems can instead reuse evidence or prior experience at inference time. However, persistent storage alone does not determine whether newly derived experience should become reusable knowledge. Feedback from an individual case may encode a transferable correction, but it may also reflect noise, redundancy, distribution-specific shortcuts, or conflicts with previously useful rules. Once admitted into persistent memory, such information can influence later cases through retrieval.

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.

This creates an inference-time stability–plasticity problem: memory must remain sufficiently plastic to absorb reusable corrections while remaining sufficiently stable to resist harmful updates. The central question is therefore not simply whether memory helps, but when feedback-driven memory improves reasoning and when it damages previously useful knowledge. We argue that reliable adaptation requires explicit control over the memory write path—what experience is admitted, how it is reconciled with existing knowledge, and how it is later reused. This differs from ordinary case-memory retrieval, which primarily reuses prior episodes, and from global context optimization, which may rewrite a shared prompt or playbook without explicitly separating stable knowledge from revisable experience.

To address this problem, we propose a feedback-driven context evolution framework that treats test-time adaptation as controlled modification of reusable external memory. The framework separates stable clinical knowledge from revisable case-derived experience so that new information can accumulate without directly overwriting trusted guidance. Case-level feedback is converted into candidate experience and passed through an auditable write process that determines whether and how it may modify active memory. A deterministic Gate validates updates, consolidates compatible rules, isolates conflicts, and withholds ambiguous candidates from active memory. During inference, section-first retrieval selectively combines relevant stable knowledge with accumulated experience according to case relevance and the configured context capacity. The write path, rather than memory persistence alone, thus becomes the primary control point for balancing adaptation and knowledge preservation.

We use complete blood count (CBC) based case streams as a constrained clinical testbed because their inputs are standardized, evidence boundaries are explicit, and clinically different conditions may exhibit overlapping laboratory patterns. We therefore study disease detection, tumor risk assessment, and ranked tumor-group prioritization under limited evidence rather than definitive diagnosis. This setting enables controlled analysis of accuracy, memory evolution, and cross-backbone reuse. We make the following contributions: (1) We propose a feedback-driven context evolution framework for test-time adaptation of frozen LLMs across sequential laboratory cases, formulating adaptation as controlled evolution of reusable reasoning context that separates stable knowledge from revisable experience while preserving update provenance. (2) We introduce an auditable memory evolution mechanism that converts case-level feedback into candidate experience linked to provenance and regulates memory admission through deterministic write gating, compatibility-aware consolidation, and conflict isolation; section-first retrieval then selectively reuses stable and dynamic knowledge during inference. (3) We construct a CBC-based case stream evaluation setting and a reliability-oriented protocol for studying the stability–plasticity trade-off of feedback-driven memory, jointly evaluating task accuracy and long-horizon evolution across predefined memory-update checkpoints, and cross-backbone memory reuse under read-only evaluation.

2 RELATED WORK

LLMs for Medical Reasoning and Clinical Memory. LLMs have been increasingly explored for medical question answering, dialogue support, and clinical reasoning (Nori et al., 2023a). Existing approaches improve medical competence through biomedical pretraining, supervised instruction tuning, or parameter-efficient adaptation (Tu et al., 2025; Hager et al., 2024; Wang et al., 2024; Qiu et al., 2025; Zhu et al., 2025). Laboratory-centered reasoning is more constrained because models must connect abnormal numeric findings with differential clues, confounders, and qualified interpretations rather than merely recall medical facts (Khandekar et al., 2024). Benchmarks for interactive EHR use and emergency-room reasoning further motivate workflow-aware evaluation (Jiang et al., 2025; Mehandru et al., 2025). Most closely, MedChain models sequential clinical workflows and introduces MedChain-Agent, whose feedback mechanism and MCase-RAG module reuse previous cases (Liu et al., 2025), while MemoryART combines working, episodic, and semantic memory for long-horizon healthcare dialogue (Dai et al., 2026). These systems establish the value of persistent clinical memory. Our focus is narrower: rather than retaining patient histories or resolved cases as demonstrations, we study how audited feedback can become reusable cross-case reasoning rules and how admission of such rules into active memory should be controlled.

Context Engineering and Test-Time Adaptation. A complementary line of work adapts model behavior through prompts, retrieved evidence, or external memory without modifying model weights

(Wang et al., 2022; Asai et al., 2024; Nori et al., 2023b; Yang et al., 2025; Xiong et al., 2024). Recent approaches differ mainly in what they store, retrieve, and update. Retrieval-oriented systems organize prior experience: SECOM constructs segment-level memory with compression-based denoising (Pan et al., 2025), MemTree organizes information into hierarchical abstractions (Rezazadeh et al., 2025), and Agent S retrieves narrative and episodic experience for hierarchical planning (Agashe et al., 2025). Evolving-memory methods move beyond static retrieval: A-MEM creates interconnected structured notes and updates related historical memories as new memories arrive (Xu et al., 2025), while SEGMEM-RAG accumulates reflective experience to adapt retrieval across changing knowledge sources (Fan et al., 2026). Context-optimization approaches such as GEPA (Agrawal et al., 2025), Dynamic Cheatsheet (Suzgun et al., 2026), and ACE (Zhang et al., 2026) refine prompts, self-curated memories, or evolving playbooks from task experience. MemoryAgentBench further emphasizes test-time learning, long-range understanding, and conflict resolution beyond retrieval accuracy (Hu et al., 2026). Together, these studies expose a stability–plasticity question: new experience should remain learnable without indiscriminately displacing useful information. Our framework addresses this at the write path by separating stable knowledge from revisable experience and controlling candidate admission before it changes active memory; provenance and capacity management support auditability rather than serving as independent learning mechanisms.

Feedback-Driven and Self-Evolving Agents. Language agents can improve through self-critique, execution feedback, tool-assisted verification, and stored experience (Madaan et al., 2023; Schick et al., 2023; Gou et al., 2024; Bai et al., 2022; Yuan et al., 2024; Chen et al., 2024; Liao et al., 2025; Tran et al., 2025). Reflexion, Self-Refine, and related systems use feedback to revise trajectories or derive reusable guidance, while later approaches make adaptation persistent across tasks. ReasoningBank distills structured reasoning strategies from successful and failed trajectories and retrieves them during test-time learning (Ouyang et al., 2026). SKE-Learn verifies self-extracted knowledge to select synthetic training traces and updates model capabilities through iterative self-training (Huang et al., 2025), making it complementary to frozen-parameter memory adaptation. Dynamic Cheatsheet maintains self-curated transferable strategies, ACE uses Generator–Reflector–Curator roles with incremental playbook updates, and A-MEM evolves representations and links among related memory notes (Suzgun et al., 2026; Zhang et al., 2026; Xu et al., 2025). These methods show that feedback-derived experience can become persistent, making write quality consequential once new information influences future tasks. Our framework therefore targets whether and how audited case-derived experience should modify reusable memory while stable clinical knowledge remains protected.

3 METHODS

3.1 Problem Setting and Framework Overview

We study feedback-driven adaptation for frozen LLMs over sequential CBC cases. Each case x_t contains a normalized CBC panel and, when available, limited demographic variables. Because CBC findings alone rarely support definitive diagnosis, we formulate the task as risk-oriented reasoning under bounded evidence. The model determines disease status, assesses tumor-related risk when applicable, and ranks candidate tumor groups for tumor-oriented differentiation. Throughout adaptation, the backbone LLM remains frozen; only the external reasoning context evolves.

Our objective is to balance *plasticity*, allowing useful corrections from prior cases to influence future reasoning, with *stability*, protecting read-only prior guidance from feedback-driven modification from noisy or conflicting feedback. As illustrated in Fig. 1, the framework contains three coupled mechanisms: a hierarchical memory separating stable knowledge from revisable experience, a deterministic write gate controlling whether feedback-derived candidates may modify active memory, and section-first retrieval constructing case-specific context from relevant stable and dynamic information. Together, they implement controlled external-memory adaptation rather than parameter learning or unrestricted case accumulation.

3.2 Hierarchical Memory for Stable Knowledge and Revisable Experience

Memory provides the persistent external substrate for preserving stable knowledge and accumulating auditable case-derived experience. At case step t, we represent the persistent memory as

$$\mathcal{M}_t = \big(\mathcal{K}, \mathcal{E}_t, \mathcal{I}_t\big), \tag{1}$$

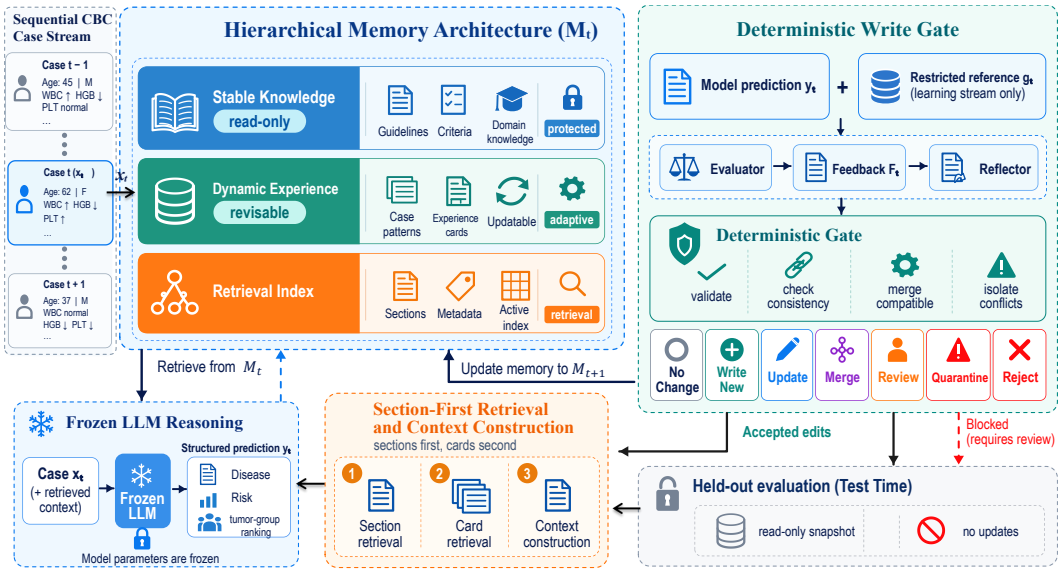


Figure 1: Overview of the feedback-driven context evolution framework. For each incoming CBC case, the framework retrieves structured medical memory and supplies it to the frozen LLM reasoner to generate a structured prediction conditioned on the CBC input and retrieved memory. The prediction is evaluated against learning-case reference feedback, and admitted feedback-derived edits update the memory for future cases.

where $\mathcal{K}$ denotes stable knowledge, $\mathcal{E}_t$ denotes dynamic experience accumulated before case t, and $\mathcal{I}_t$ denotes the retrieval catalog. $\mathcal{K}$ contains pre-adaptation guidance and training-derived summaries that remain read-only during a case stream. Stable denotes update status rather than verified medical correctness. In contrast, $\mathcal{E}_t$ contains provisional, reusable lessons derived from audited learning cases and may be inserted, consolidated, deactivated, or isolated without modifying $\mathcal{K}$. This separation keeps new experience editable while preventing individual corrections from overwriting stable knowledge. Throughout the paper, stable denotes read-only status during adaptation rather than independently verified medical correctness, and auditable denotes provenance-linked inspectability rather than clinical validation.

The memory has three logical levels. The audit level stores immutable case events, including model output, applicable feedback, the derived candidate, and the archive decision. The card level stores concise, reasoning units visible to the model. A dynamic card e_i records a semantic section s_i, a reusable reasoning rule r_i, its applicability a_i and limitations ℓ_i, provenance p_i, and version/status metadata v_i:

$$e_i = (s_i, r_i, a_i, \ell_i, p_i, v_i), \tag{2}$$

Stable knowledge uses the same card interface but remains read-only. The catalog level stores compact section descriptors and active-card references, enabling retrieval over a structured index rather than the complete history.

Case-derived experience is represented as generalized reasoning guidance rather than stored patient examples. Before becoming model-visible, a candidate must satisfy structural validation over required fields, allowable scope, provenance, and content form. Case-specific identifiers, copied patient passages, and rules tied to an individual numerical instance are excluded. Structural validation establishes eligibility for the subsequent write-path checks. Reused dynamic cards remain linked to the audit events that created or revised them, preserving traceability between feedback, memory evolution, and later predictions.

Stable cards are protected from feedback-driven replacement, whereas inactive dynamic cards can be removed from active retrieval without deleting their provenance. Compatible dynamic cards may be consolidated while preserving source links and version history. Versioned memory states separate learning from held-out evaluation: evaluation loads a fixed read-only state, preventing test cases from entering the write path.

3.3 FEEDBACK-GOVERNED MEMORY WRITING

For a learning case x_t, the framework first constructs a retrieved context C_t and obtains a structured prediction

$$y_t = f_{\text{diag}}(x_t, C_t), \tag{3}$$

where f_{diag} denotes the frozen LLM reasoning component. A reference g_t is available only within the designated learning stream and is never included in held-out retrieval. The evaluator compares y_t with the applicable reference and produces hierarchical feedback

$$F_t = \left(f_t^{\text{det}}, f_t^{\text{risk}}, f_t^{\text{rank}}\right), \tag{4}$$

where f_t^{det}, f_t^{risk}, and f_t^{rank} correspond to disease detection, tumor risk assessment, and candidate tumor group ranking, respectively. Inapplicable components remain undefined. A reflector then transforms the prediction and applicable feedback into either no proposed change or a generalized candidate experience $\tilde{e}_t$.

Reflection does not directly update memory. Each candidate passes through a deterministic write gate

$$G\big(\tilde{e}_t; \mathcal{M}_t\big), \tag{5}$$

which determines whether and how it may affect active experience. The Gate checks structural admissibility, task and section compatibility, internal consistency, and its relation to active experience. Compatible candidates may extend or merge with existing rules while preserving provenance. Unresolved contradictions are isolated from active retrieval, ambiguous cases are withheld for human review, and invalid or policy-disallowed candidates are rejected.

The resulting semantic memory transition is

$$\mathcal{M}_{t+1} = \mathcal{U}\left(\mathcal{M}_t, G(\tilde{e}_t; \mathcal{M}_t)\right), \tag{6}$$

where $\mathcal{U}$ applies an admitted edit to the active rule content; when no semantic edit is admitted, the active rule content remains unchanged. This makes the write path the primary control point of adaptation. Unlike ordinary episodic memory, which mainly retains previous examples, the Gate separates candidate generation from admission into reusable memory. Plasticity therefore arises from feedback-derived candidates, whereas stability is enforced by protecting $\mathcal{K}$ and constraining changes to $\mathcal{E}_t$.

References are used only for designated learning cases. Held-out cases use the same frozen backbone and retrieval process with memory writing disabled, so their labels cannot affect current or subsequent predictions. Versioned snapshots allow each prediction to be attributed to the memory state used to construct its context.

3.4 SECTION-FIRST RETRIEVAL AND CONTROLLED CONTEXT EVOLUTION

The third mechanism determines how stable knowledge and accumulated experience are reused. Direct retrieval from a flat memory can allow dense or recently updated regions to dominate context construction. We therefore use section-first retrieval, separating coarse semantic selection from card-level selection.

For a model-visible case x_t, the retrieval query contains only available case fields and the task instruction; reference labels, source identities, patient group identifiers, and evaluation metadata are excluded. The retriever first ranks semantic sections in $\mathcal{I}_t$, then ranks active cards within the selected sections to form an ordered candidate set $\mathcal{B}_t$. The reasoning context is constructed as

$$C_t = \text{Prompt}\left(x_t, \text{Pack}\big(\mathcal{B}_t; \mathcal{M}_t\big)\right), \tag{7}$$

where Pack selects complete memory cards that fit the configured context capacity.

Section-first retrieval constrains competition to predefined semantic memory regions before card-level selection and makes context composition explicit. Stable and dynamic cards are selected according to case relevance within configuration-specific caps, so their realized number and mixture can vary across cases while the context remains bounded. When the candidate set exceeds capacity, lower-ranked cards are omitted instead of partially truncating a rule or patient input. The framework records selected sections, retrieved card identifiers, their stable/dynamic status, and provenance links, associating each prediction with the memory state and experience used for reasoning.

The complete adaptation cycle is therefore

$$\mathcal{M}_t \xrightarrow{\text{retrieve}} C_t \xrightarrow{\text{reason}} y_t \xrightarrow{\text{audit}} F_t \xrightarrow{\text{reflect}} \tilde{e}_t \xrightarrow{\text{gate/update}} \mathcal{M}_{t+1}. \tag{8}$$

At each learning step, retrieval constructs a case-specific context, the frozen LLM produces a structured prediction, the evaluator produces feedback, and the reflector converts applicable feedback into either no proposed change or a candidate experience. The Gate then determines whether and how that candidate may modify future memory. The distinguishing design choice is the separation of candidate generation from controlled admission into reusable memory. Our framework evolves a provenance-preserving collection of reusable rules under explicit write control while retaining stable knowledge as a protected substrate. During held-out evaluation, the cycle terminates after reasoning and the loaded memory snapshot remains unchanged.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETTINGS

Dataset and Evaluation Protocols. We evaluate CBC-based risk-oriented reasoning using de-identified hospital records and MIMIC-IV (Johnson et al., 2023), covering normal, non-tumor diseases, and eight tumor groups. The knowledge package contains 192 memory items, including 20 statistical anchors estimated from training-only data; their construction and selection are detailed in Appendix A.3. *Mixed-Source Adaptation* (MSA) comprises 350 initialization, 700 update, and 780 test cases, with hospital and MIMIC cases mixed during initialization and updating. *Hospital-to-MIMIC Transfer* (HMT) comprises 700 hospital initialization cases, 700 MIMIC update cases, and 1,200 test cases, examining memory adaptation across a source transition. Within each protocol, patients are disjoint across phases, and test cases are evaluated with memory writing disabled. The test sets contain 180/90/510 normal/non-tumor/tumor cases for MSA and 400/400/400 for HMT, with equal hospital and MIMIC totals. Appendix A details the splits and label space. MSA and HMT are alternative partitions of shared source pools rather than independent cohorts. All normal controls are hospital-derived, whereas MIMIC negatives are inpatient admissions without the specified malignancy codes in their available diagnosis history.

Tasks and Metrics. We evaluate disease detection (S1), tumor risk assessment among reference-positive disease cases (S2), and tumor group ranking among reference-positive tumor cases (S3). Table 1 reports accuracy (Acc), macro-F1 (Ma-F1), and support-weighted F1 (Wt-F1). For S3, Acc measures whether the highest-ranked candidate matches the reference tumor group, while Acc@3 measures whether the reference group appears in the emitted shortlist of three candidates. For each stage, Acc and Acc@3 are computed over the corresponding reference-eligible population, including upstream routing misses. Response coverage (Cov) separately measures the fraction of the same reference-eligible population for which a valid stage-specific output is available. The reference-eligible S1/S2/S3 populations are fixed at 780/600/510 for MSA and 1,200/800/400 for HMT. Appendix D provides the scoring definitions.

Backbone and Comparison Methods. Qwen3.8-27B (Qwen Team, 2026) is the principal frozen backbone. We compare our framework with a medically prompted Frozen_LLM, static RAG (Lewis et al., 2020), Dynamic Cheatsheet (DC) (Suzgun et al., 2026), GEPA (Agrawal et al., 2025), and ACE (Zhang et al., 2026). Together, these comparisons cover fixed prompting, retrieval, and context optimization. The memory evolution protocol evaluates snapshots after 0, 100, 300, 500, and 700 update cases. Configuration comparisons vary Gate, context capacity, and statistical anchors. Appendix C records implementation details; Appendix D.3 specifies the statistical reporting and verification protocol. All differences below are descriptive percentage-point differences between the displayed values.

4.2 MAIN RESULTS

Normal-versus-disease classification. To assess whether controlled memory evolution improves disease recognition beyond static retrieval and adaptive context methods, we compare all methods in Table 1. On the constructed MSA and HMT protocols, our framework has the highest displayed S1 normal-versus-disease accuracy on MSA (87.43%) and HMT (82.92%). Compared with the

Table 1: Reported task performance on MSA and HMT (%). Bold indicates the largest displayed value within each protocol and metric. Aggregation and provenance checks are specified in Appendix D.3.

Protocol	Method	Disease detection			Tumor risk			Tumor group ranking			
		Acc	Ma-F1	Wt-F1	Acc	Ma-F1	Wt-F1	Acc	Acc@3	Ma-F1	Wt-F1
MSA	Frozen_LLM	76.92	75.20	79.05	29.17	30.70	39.47	12.35	18.04	9.33	9.33
MSA	RAG	85.77	85.23	87.70	**73.17**	43.90	74.63	22.55	43.53	15.42	15.43
MSA	DC	64.62	67.85	69.08	37.67	33.40	51.45	15.88	24.12	11.28	11.32
MSA	GEPA	81.15	78.48	82.56	59.33	48.25	68.65	19.80	37.45	13.82	13.83
MSA	ACE	86.41	83.74	87.29	58.67	49.58	67.74	18.24	35.10	13.31	13.31
MSA	Ours	**87.43**	**85.36**	**88.47**	70.53	**51.49**	**74.99**	**22.63**	**43.55**	**15.73**	**15.73**
HMT	Frozen_LLM	70.83	71.68	71.79	26.25	33.18	33.18	10.25	13.25	9.09	9.09
HMT	RAG	78.83	79.93	80.33	42.63	36.59	36.59	18.75	42.50	13.29	13.29
HMT	DC	60.08	63.06	61.13	26.88	35.56	35.56	15.75	23.50	12.58	12.58
HMT	GEPA	81.83	81.40	82.35	46.50	52.78	52.78	18.50	40.50	**15.89**	**15.89**
HMT	ACE	80.00	79.64	80.54	49.50	**54.83**	**54.83**	20.25	42.00	15.22	15.22
HMT	Ours	**82.92**	**82.45**	**83.41**	**50.06**	52.05	52.05	**21.20**	**44.41**	14.01	14.01

strongest context-based competitors, ACE on MSA and GEPA on HMT, the gains are 1.02 and 1.09 points; compared with Frozen LLM, they are 10.51 and 12.09 points. Our method also gives the highest S1 macro-F1 and weighted F1 in both protocols. Thus, the improvement remains visible against RAG and context optimization baselines, although the margin over the strongest adaptive methods is smaller.

Non-Tumor vs. Tumor-Related Classification and Candidate Ranking. To test whether this advantage extends to later stages, we compare the same methods on S2 and S3. The S2 result is metric-dependent: on MSA, RAG has higher accuracy than ours (73.17% vs. 70.53%), whereas our method gives the highest macro-F1 and weighted F1; on HMT, our method gives the highest accuracy (50.06%), while ACE and GEPA remain stronger on F1. S3 shows a clearer pattern: our method gives the highest top-1 accuracy and Acc@3 in both protocols. On MSA, however, its margins over RAG are only 0.08 and 0.02 points, indicating a near tie; the HMT advantage is larger. Overall, feedback-driven memory is most consistent for disease detection and candidate prioritization, while tumor-risk discrimination remains metric- and protocol-dependent.

Long-Horizon Memory Evolution. To examine how performance changes as feedback-derived experience accumulates, Figure 2 evaluates a fixed test cohort at five memory snapshots. The trajectories increase across all three stages in both protocols, with HMT showing a larger late-stage change in tumor-risk accuracy and Acc@3 and MSA a more gradual progression. The curves therefore show that performance improves progressively across checkpoints as the number of processed update cases increases.

Cross-Backbone Memory Reuse. To examine whether learned memory can be consumed by other frozen backbones, Table 2 reuses Qwen-derived memory with Gemma 4 26B (Google DeepMind, 2026) and MedGemma-27B (Sellergren et al., 2025). Both recipients improve S1 accuracy and MSA S2 accuracy relative to Qwen, but HMT S2 and S3 scores generally decrease. The results demonstrate cross-backbone compatibility rather than a universal transfer gain; matched recipient baselines are needed to isolate the benefit of reuse.

4.3 Ablation Study

Table 3 examines controlled memory admission, machine-learning statistical guidance, and retrieval-context allocation. Under the larger-context setting without anchors, the Gate-enabled configuration shows higher S1, S2, and S3 accuracy by 4.40, 3.13, and 4.09 points, respectively, together with a 0.34-point higher S3 Acc@3. With Gate enabled, the anchor-enabled configuration likewise shows higher S1, S2, and S3 accuracy by 1.68, 5.51, and 1.31 points and a 2.45-point higher Acc@3 than the anchor-disabled configuration. The largest Gate-associated difference occurs in disease

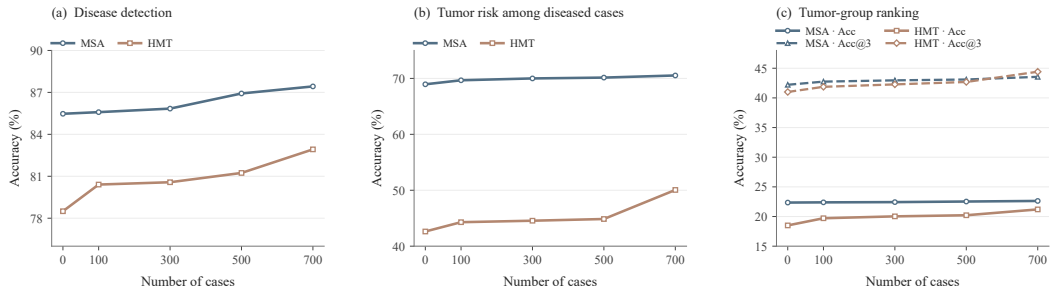


Figure 2: Accuracy across memory-update checkpoints on MSA and HMT. Panels show (a) disease detection, (b) tumor risk assessment, and (c) tumor group ranking with both top-1 accuracy (Acc, solid lines) and Acc@3 (dashed lines). The horizontal axis counts processed update cases; each protocol uses a fixed test cohort. Panels use separate vertical scales. Statistical reporting details are specified in Appendix D.3.

Table 2: Cross-backbone memory reuse (%). Qwen3.8-27B is the source-model reference and reproduces Ours from Table 1; Gemma 4 26B and MedGemma-27B are the recipient models. Differences describe performance across backbones using the memory-reuse setting.

Protocol	Backbone	Disease detection			Tumor risk			Tumor group ranking			
		Acc	Ma-F1	Wt-F1	Acc	Ma-F1	Wt-F1	Acc	Acc@3	Ma-F1	Wt-F1
MSA	Qwen3.8-27B	87.43	85.36	88.47	70.53	51.49	74.99	22.63	43.55	15.73	15.73
MSA	Gemma 4 26B	92.44	90.29	92.75	80.17	45.64	77.58	19.61	43.73	11.51	11.56
MSA	MedGemma-27B	92.69	91.17	93.28	80.67	45.57	77.48	14.71	39.02	7.73	7.75
HMT	Qwen3.8-27B	82.92	82.45	83.41	50.06	52.05	52.05	21.20	44.41	14.01	14.01
HMT	Gemma 4 26B	84.17	83.44	84.60	47.25	38.19	38.19	18.00	41.00	10.76	10.76
HMT	MedGemma-27B	87.25	85.43	87.16	49.63	34.89	34.89	13.00	38.00	4.59	4.59

detection, whereas the largest anchor-associated difference occurs in tumor risk assessment, indicating that the observed configuration differences are concentrated at different decision stages. These comparisons should be interpreted at the configuration level because anchor-disabled runs rebuild memory from initialization and the admission policies retain their associated consolidation behavior.

Context and Retrieval Capacity. Without Gate or anchors, the larger-context configuration shows higher S1/S2/S3 accuracy by 2.19/1.62/1.05 points and a 6.37-point higher S3 Acc@3 than the smaller-context configuration. Because these profiles jointly differ in context window and stable-card capacity, the larger shortlist difference indicates sensitivity to the combined retrieval capacity setting rather than an isolated context window effect. Appendix E.2 records the paired configuration details required for component-level interpretation.

Table 3: Reported configuration comparison on HMT (%). Bold indicates the largest displayed value per metric. Configuration and pairing details are specified in Appendix E.2.

Context	Gate	Statistical anchors	S1 Acc	S2 Acc	S3 Acc	S3 Acc@3
16K	Yes	Yes	**82.92**	**50.06**	**21.20**	**44.41**
16K	Yes	No	81.24	44.55	19.89	41.96
16K	No	No	76.84	41.42	15.80	41.62
4K	No	No	74.65	39.80	14.75	35.25

4.4 ERROR STRUCTURE AND RESPONSE COVERAGE

Class-Balanced Recognition and Response Coverage. To characterize class balance and stage-level output availability, Table 4 reports Bal.Acc and Cov. Our framework gives the highest disease-detection balanced accuracy on MSA and HMT (91.39% and 87.13%) and the highest HMT tumor-

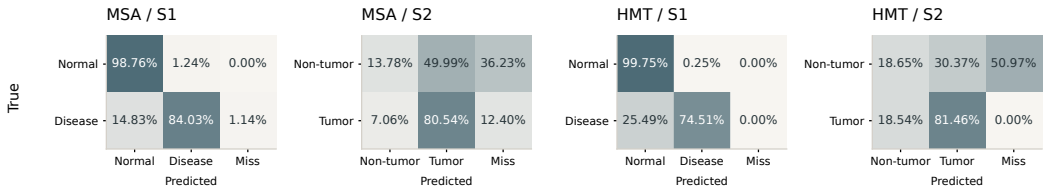


Figure 3: Stage-wise confusion matrices for MSA and HMT. Rows indicate reference classes and columns indicate predicted classes, with the supplied percentages normalized within each row. S1 distinguishes normal from disease; S2 distinguishes non-tumor disease from tumor. The Miss column records the figure's missing-output category. Scoring and routing verification are documented in Appendix E.

risk balanced accuracy. Tumor-risk coverage reaches 84.03% and 74.51%, exceeding the strongest comparison values by 1.53 and 1.63 points. However, RAG retains higher tumor-group coverage in both protocols, and ACE slightly exceeds our MSA tumor-risk balanced accuracy. The results support improved class-balanced disease recognition without uniform superiority.

Stage-Wise Prediction Patterns. Figure 3 localizes the remaining errors. At S1, normal-class recall is 98.76% on MSA and 99.75% on HMT, while disease recall is 84.03% and 74.51%. At S2, tumor recall remains high (80.54% and 81.46%), whereas non-tumor recall is much lower, with many non-tumor cases assigned to the tumor or missing-output columns. This asymmetry explains the metric-dependent S2 results and highlights non-tumor confounder discrimination as a remaining challenge.

Table 4: Class-balanced recognition and response coverage (%). Bold indicates the largest displayed value within each protocol and metric, including ties. Statistical reporting follows Appendix D.3.

Protocol	Method	Disease detection		Tumor risk		Tumor group ranking	
		Bal.Acc	Cov	Bal.Acc	Cov	Bal.Acc	Cov
MSA	Frozen_LLM	85.00	98.59	32.25	70.00	12.32	27.84
MSA	RAG	90.56	96.79	43.04	81.67	22.58	**86.08**
MSA	DC	77.00	88.85	25.36	54.00	15.82	42.94
MSA	GEPA	87.56	**100.00**	42.68	75.67	19.83	66.47
MSA	ACE	90.97	**100.00**	**47.32**	82.50	18.24	63.53
MSA	Ours	**91.39**	99.13	47.16	**84.03**	**22.67**	80.54
HMT	Frozen_LLM	78.13	97.58	26.25	56.25	10.25	19.50
HMT	RAG	84.13	96.83	42.63	68.25	18.75	**84.50**
HMT	DC	70.06	90.25	26.88	40.13	15.75	44.50
HMT	GEPA	86.31	**100.00**	46.50	72.88	18.50	61.25
HMT	ACE	85.00	**100.00**	49.50	70.00	20.25	73.50
HMT	Ours	**87.13**	**100.00**	**50.06**	**74.51**	**21.20**	81.46

5 CONCLUSION

We present an auditable feedback-driven context evolution framework that adapts frozen LLMs through protected stable knowledge, gated case-derived experience, and section-first retrieval. Across MSA and HMT, the framework most consistently improves disease detection and candidate tumor group ranking relative to retrieval and context optimization baselines, whereas tumor risk assessment remains metric- and protocol-dependent. Checkpoint, ablation, and cross-backbone analyses further show sensitivity to admission configuration, retrieval configuration, and recipient backbone. Overall, the results support controlled memory evolution as a practical way to study the stability-plasticity trade-off in CBC-based risk-oriented reasoning rather than as a universally superior replacement for simpler retrieval. Limitations and directions for future work are discussed in Appendix G.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and language editing, and were also used as development aids for selected research tasks such as code debugging, literature organization, and result interpretation. All experimental configurations, implementations, reported results, and scientific conclusions were reviewed and verified by the authors, who take responsibility for the final manuscript.

ETHICS STATEMENT

Use of the retrospective hospital dataset was approved by the relevant institutional ethics committee under the applicable protocol. MIMIC-IV was accessed by credentialed researchers under its data-use requirements. All experiments involving these datasets were conducted on locally deployed machines. Study data remained within the local computing environment and were not transmitted to external services or disclosed to unauthorized parties.

REPRODUCIBILITY STATEMENT

For experimental details, we refer readers to Appendix A for dataset construction and knowledge resources, Appendix B for the memory workflow, Appendix C for implementation and prompt configurations, and Appendices D and E for evaluation metrics, statistical reporting, and additional experimental documentation. The source code will be released as open source upon acceptance of the paper.

REFERENCES

Saaket Agashe, Jiuzhou Han, Shuyu Gan, Jiachen Yang, Ang Li, and Xin Eric Wang. Agent s: An open agentic framework that uses computers like a human. In *International Conference on Learning Representations*, 2025.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2310.11511.

Yuntao Bai, Saurav Kadavath, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint*, 2022. URL https://arxiv.org/abs/2212.08073.

Balu Bhasuran, Qiao Jin, Yuzhang Xie, Carl Yang, Karim Hanna, Jennifer Costa, Cindy Shavor, Wenshan Han, Zhiyong Lu, and Zhe He. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. *npj Digital Medicine*, 8(1):166, 2025.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2401.01335.

Renke Dai, Hebin Hu, Jiahui Zhang, Yilin Kang, and Ah-Hwee Tan. Memoryart: Enhancing llms via multi-memory models with adaptive resonance theory for healthcare agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026.

Xuanbo Fan, Tianqi Zhao, Yi Cheng, Chi Xiu, Jiaxin Guo, Boci Peng, Bingjing Xu, Jessica Zhang, Feng Sun, and Yan Zhang. Segmem-rag: Adaptive memory for retrieval-augmented generation in open-ended knowledge environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026.

Google DeepMind. Gemma 4 Model Card. https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Accessed 2026-09-20; 26B A4B variant.

Zhibin Gou, Zhihong Shao, et al. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2305.11738.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Raphael Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision making. *Nature Medicine*, 30(9):2600–2610, 2024. doi: 10.1038/s41591-024-03097-1. URL https://www.nature.com/articles/s41591-024-03097-1.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026.

Shijue Huang, Wanjun Zhong, Deng Cai, Fanqi Wan, Chengyi Wang, Mingxuan Wang, Mu Qiao, and Ruifeng Xu. Empowering self-learning of llms: Inner knowledge explicitation as a catalyst. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2501.14654.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S. Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W. Safranek, Abid A. Anwar, Andrew Zhang, Aidan Gilson, Maxwell B. Singer, Amisha Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, and Zhiyong Lu. Medcalc-bench: Evaluating large language models for medical calculations. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2406.12036.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. Reflectool: Towards reflection-aware tool-augmented clinical agents. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2410.17657.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael R. Lyu, and Wenting Chen. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2303.17651.

Liam G McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, Stephen Bacchi, Jie Ming Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A Buckley, Peter Brodeur, et al. Assessment of large language models in clinical reasoning: a novel benchmarking study. *NEJM AI*, 2(10):AIdbp2500120, 2025.

Nikita Mehandru, Niloufar Golchini, Namrata Garg, et al. Er-reason: A benchmark dataset for llm clinical reasoning in the emergency room. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2505.22919.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023a.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023b.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. On memory construction and retrieval for personalized conversational agents. In *International Conference on Learning Representations*, 2025.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, et al. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799, 2025. doi: 10.1038/s41467-025-64769-1. URL https://www.nature.com/articles/s41467-025-64769-1.

Qwen Team. Qwen3.8-27B: Model card. https://huggingface.co/Qwen/Qwen3.8-27B, 2026.

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms. In *International Conference on Learning Representations*, 2025.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2302.04761.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, et al. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*, 2025. doi: 10.48550/arXiv.2507.05201. URL https://arxiv.org/abs/2507.05201.

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.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7080–7106, 2026.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Hieu Tran, Zonghai Yao, Zhichao Yang, Junda Wang, Yifan Zhang, Shuo Han, Feiyun Ouyang, and Hong Yu. Rare: Retrieval-augmented reasoning enhancement for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18305–18330, 2025. doi: 10.18653/v1/2025.acl-long.896. URL https://aclanthology.org/2025.acl-long.896/.

Tao Tu, Mike Schaekermann, Anil Palepu, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025. doi: 10.1038/s41586-025-08866-7. URL https://www.nature.com/articles/s41586-025-08866-7.

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*, 2022. doi: 10.48550/arXiv.2203.11171. URL https://arxiv.org/abs/2203.11171.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, et al. ConU: Conformal uncertainty in large language models with correctness coverage guarantees. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2407.00499.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2402.13178.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Rui Yang, Yilin Ning, Emilia Keppo, et al. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):2, 2025. doi: 10.1038/s44401-024-00004-1. URL https://www.nature.com/articles/s44401-024-00004-1.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2401.10020.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, et al. Diagnosisarena: Benchmarking diagnostic reasoning for large language models. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2505.14107.

APPENDIX

CONTENTS

A Dataset Construction and Knowledge Resources 16
A.1 Patient-Level Partitions and Source Composition . . . 16
A.2 Tumor Label Space and Input Processing . . . 17
A.3 Knowledge Initialization and Statistical Anchors . . . 18
B Memory Representation and Operational Workflow 19
B.1 Memory Card and Audit Schemas . . . 19
B.2 Update and Evaluation Procedures . . . 20
B.3 Extended Feedback Decomposition . . . 20
B.4 Write Decisions and Active-Memory Semantics . . . 21
B.5 Section Retrieval and Context Packing . . . 22
B.6 Recursive Evolution and Read-Only Evaluation . . . 22
C Implementation Details and Prompt Templates 23
C.1 Backbone, Context, and Decoding Configuration . . . 23
C.2 Baseline Adaptations and Matched Budgets . . . 24
C.3 Exact Prompt Documentation . . . 24
D Evaluation Metrics and Statistical Reporting 28
D.1 Reference-Defined Stage Eligibility . . . 28
D.2 Accuracy, F1, Balanced Accuracy, and Coverage . . . 29
D.3 Checkpoint Aggregation and Paired Comparisons . . . 30
D.4 Claim-Level Faithfulness: Reserved Evaluation . . . 30
D.5 Conditional Count-Consistency Checks . . . 31
E Supplementary Results and Experimental Details 31
E.1 Complete Endpoint and Longitudinal Results . . . 31
E.2 Matched Component Comparisons . . . 32
E.3 Cross-Backbone Transfer Protocol . . . 33
E.4 Confusion Matrices, Subgroups, and Reliability . . . 33
E.5 Computational Cost and Memory Growth . . . 34
F Qualitative Cases and Memory Audit Trails 34
F.1 Simulated Question Answering and Result Display . . . 35
F.2 Prediction Context and Memory Provenance . . . 37
G Limitations and Future Work 37
G.1 Limitations . . . 38
G.2 Future Work . . . 38

A Dataset Construction and Knowledge Resources

Return to appendix contents

A.1 Patient-Level Partitions and Source Composition

Source cohorts and reference labels. The evaluation uses de-identified hospital records and a patient-level cohort derived from MIMIC-IV 3.1 (Johnson et al., 2023). The hospital source contains 5,288 records: 1,322 tumor cases, 2,644 non-tumor disease cases, and 1,322 normal controls. Its cohort documentation identifies the CBC as the first admission examination and the disease reference as the subsequent final discharge diagnosis; normal status is inherited from the source control definition. Complete examination-to-diagnosis intervals are not available for this cohort. We therefore use these records for retrospective classification rather than prediction at a uniform time horizon.

For MIMIC, the index panel is the earliest specimen with all nine core CBC measurements available within 0–24 hours after admission. Measurements are grouped by specimen, rather than assembled from unrelated sampling times. Each patient contributes the first eligible admission, ordered by admission time and admission identifier. A positive reference requires a malignant-neoplasm code in the same admission's final diagnosis records (ICD-10 C00–C97 or ICD-9 140–208). A negative reference requires the absence of such codes throughout that patient's available MIMIC diagnosis history. This latter definition concerns the observed database history, not lifelong absence of malignancy. These discharge-code references are distinct from pathological confirmation.

The experiments draw MIMIC cases from the cancer-to-non-cancer ratio-matched source cohort, comprising 25,544 positive and 76,632 inpatient negative cases. Of the positive cases, 9,493 map to the eight-group label space in Appendix A.2; the remaining 16,051 are excluded rather than reassigned to a target group. The subsequent protocol-specific sampling defines the evaluation class proportions.

Protocol construction and isolation. Memory initialization, sequential updating, and held-out evaluation use separate patient groups within each protocol. The sampling procedure reserves test patients first, then selects initialization and update cases without reusing their patient-group identifiers. Selection is deterministic under the recorded sampling seeds. Table 5 summarizes the case partitions; the 192 knowledge records described in Appendix A.3 are additional initialization resources, not patients. A checkpoint of zero updates follows initialization-case learning and therefore need not have an empty dynamic memory.

Table 5: Case partitions by protocol, source, and reference category. Source totals and category totals each sum to the final column. Knowledge records are excluded.

Protocol	Phase	Hospital	MIMIC	Normal	Non-tumor	Tumor	Total
MSA	Initialization	245	105	100	100	150	350
MSA	Update	490	210	200	200	300	700
MSA	Test	390	390	180	90	510	780
HMT	Initialization	700	0	140	280	280	700
HMT	Update	0	700	0	350	350	700
HMT	Test	600	600	400	400	400	1,200

The phase manifests retain pseudonymous patient-group and case identifiers, source lineage, reference categories, and record hashes. Within each protocol, all three pairwise phase intersections are zero for both patient-group identifiers and exact question text; each phase also contains unique patient groups. This verifies isolation under the supplied identity mapping. MSA and HMT are alternative partitions of shared source pools and may reuse patients across protocols; they are not independent cohorts that can be pooled to increase sample size. At evaluation, labels are stored separately from model inputs, and the designated memory snapshot is read-only.

Equal hospital and MIMIC test totals do not imply matched source composition within each category (Table 6). All normal controls originate from the hospital; MIMIC negatives are inpatients without the specified malignancy codes, not healthy controls. HMT initializes case-derived memory on

Table 6: Test composition by source. Age is median [25th, 75th percentile] in years, computed from the model-visible age field; quartiles use linear interpolation. Female denotes the recorded sex field.

Protocol	Source	Normal	Non-tumor	Tumor	Female	Age
MSA	Hospital	180	31	179	164/390	59 [40, 71]
MSA	MIMIC	0	59	331	202/390	64 [53, 75]
HMT	Hospital	400	100	100	270/600	51 [36, 64.25]
HMT	MIMIC	0	300	300	318/600	62.5 [48, 77]

hospital records, updates it on MIMIC cases without normal controls, and evaluates it on a mixed-source test set. It consequently measures adaptation across a source transition rather than MIMIC-only external testing. The reference-eligible S1/S2/S3 test denominators are 780/600/510 for MSA and 1,200/800/400 for HMT.

A.2 Tumor Label Space and Input Processing

Closed-set label mapping. Table 7 gives the implemented crosswalk and exact test supports. For MIMIC admissions with multiple malignant diagnoses, the upstream extractor selects the malignant code with the smallest diagnosis sequence number; the partitioning stage maps its recorded family to one target group. It does not search subsequent diagnoses to rescue an otherwise excluded group. The ICD-10 and ICD-9 mappings are stored as higher- and intermediate-confidence crosswalks, respectively. These provenance categories describe mapping specificity, not expert-certified label accuracy. In particular, ICD-9 families 154, 162, and 183 can include neighboring anatomical sites, so their mappings should not be interpreted as exact equivalence to the narrower ICD-10 target families.

Table 7: Implemented tumor-group crosswalk and reference-positive tumor test counts. ICD-9 entries denote three-digit families; the ICD-10 column also defines the canonical output groups.

Tumor group	ICD-10 family	ICD-9 family	MSA	HMT
Gastric cancer	C16	151	64	50
Colorectal tumors	C18 / C20	153 / 154	63	50
Liver cancer	C22	155	64	50
Lung cancer	C34	162	64	50
Cervical cancer	C53	180	64	50
Ovarian cancer	C56	183	64	50
Prostate cancer	C61	185	63	50
Leukemia	C91 / C92	204 / 205	64	50
Total			510	400

Among the MIMIC tumor cases in the MSA test set, 156 use the ICD-10 crosswalk and 175 use the ICD-9 crosswalk; the corresponding HMT counts are 144 and 156. These counts concern the selected test cohorts, not the entire MIMIC candidate pool. Hospital references retain the source final-discharge group or normal-control designation. Diagnosis codes and mapping-confidence metadata are used for label construction and auditing, not as additional model-visible evidence.

Input fields and numerical representation. Each case is serialized as an English question containing recorded sex, age in years, and named CBC values. The case-level context field is empty. Hospital questions contain 24 CBC fields; MIMIC questions contain the 20 available fields listed in Table 8. The four hospital-only platelet indices are omitted from MIMIC questions rather than synthesized. Case keys support record alignment, while patient identifiers, source names, diagnosis codes, reference labels, and split metadata remain outside the question and retrieval query. This removes explicit source labels but does not remove source-associated measurement or missingness patterns.

The nine required MIMIC measurements are white and red blood cell counts, hemoglobin, hematocrit, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular hemoglobin con-

Table 8: CBC field inventory and MIMIC numerical scales after text conversion. Bounds are inclusive upstream data-validity filters in these scales, not clinical reference intervals. The five differential cell types are neutrophils, lymphocytes, monocytes, eosinophils, and basophils.

Field(s)	MIMIC scale	Validity bounds
White blood cell count	10^9/L	0.01–500
Five differential percentages	%	0–100
Five differential absolute counts	10^9/L	0–500
Red blood cell count	10^{12}/L	0.1–10
Hemoglobin	g/L	10–300
Hematocrit	%	1–80
Mean corpuscular volume	fL	20–200
Mean corpuscular hemoglobin	pg	5–80
Mean corpuscular hemoglobin concentration	g/L	100–600
Red cell distribution width, CV	%	5–50
Red cell distribution width, SD	fL	10–200
Platelet count	10^9/L	0–3,000
Platelet distribution width; mean platelet volume; plateletcrit; large platelet ratio	Hospital only	

centration, red cell distribution width CV, and platelet count. Missing extended measurements remain explicit as `Not available`; they are neither mean-imputed nor replaced with zero. Hospital CBC values are taken from the canonical source table. The text-conversion layer formats decimal values and the recorded sex field without fitting a scaler, imputing values, or applying additional outlier clipping. It does not append patient-specific reference intervals or unit strings to the CBC fields. Thus, normalization here denotes field and numerical-scale harmonization, not test-fitted statistical standardization.

A.3 Knowledge Initialization and Statistical Anchors

Resource composition and provenance. The stable knowledge package contains 192 items: 163 retained historical items, nine public-evidence statements, and 20 training-derived statistical anchors. The historical source contains 301 records. Grouping records by identical question text identifies 20 groups with differing reference-answer strings; all 138 records in those groups are removed, leaving 163 items. This is a reproducible text-level exclusion rule, not an adjudication that all removed variants are medically contradictory or all retained items are medically validated. The nine public statements retain source identifiers and scope restrictions, covering diagnostic and screening boundaries, biomarker limitations, and the interpretation of prognostic associations.

The initialization code converts each knowledge item deterministically into a stable card without an LLM call. The item identifier links a source event, its source line and record hash, and its card; identifiers follow the forms `L0-K-<record-id>` and `CARD-K-<record-id>`. The record identifiers distinguish historical items (`LEGACY-KNOWLEDGE-*`), public statements (`PUBLIC-CLAIM-001` through `009`), and anchors (`ML-COMMON-01` through `20`). Stable cards retain the reference text, applicability, limitations, and source-status metadata and remain read-only during sequential updating. Historical cards explicitly retain the qualification that their external verification is incomplete. Source-event text is an audit resource and is not expanded into raw patient episodes during retrieval.

Estimation cohort and exclusion boundary. The statistical anchors use a separate hospital-derived cancer-versus-non-cancer training table, restricted to records designated as training data in that source. They are not estimated from the 350 MSA or 700 HMT initialization cases. Starting from 3,702 training rows, the construction removes 1,433 rows matching a union of hospital CBC fingerprints from the initialization, update, and test partitions of both protocols. The retained cohort comprises 2,269 rows: 454 cancer and 1,815 non-cancer records.

Estimation and inclusion rule. One descriptive anchor is retained for each of the 20 shared CBC fields. For feature j, let $\bar{x}_{1j}, \bar{x}_{0j}$ and s^2_{1j}, s^2_{0j} denote the group means and sample variances among

finite observations in the cancer and non-cancer groups. The stored standardized mean difference is

$$d_j = \frac{\bar{x}_{1j} - \bar{x}_{0j}}{s_{p,j}}, \qquad s^2_{p,j} = \frac{(n_{1j} - 1)s^2_{1j} + (n_{0j} - 1)s^2_{0j}}{n_{1j} + n_{0j} - 2}. \tag{9}$$

All 20 fields have 454 cancer and 1,815 non-cancer observations in the retained table. Uncertainty is summarized by 500 within-group bootstrap resamples, recomputing both the mean difference and pooled standard deviation in every resample. The 2.5th and 97.5th percentiles form the stored interval; the NumPy random generator uses seed 82706 and advances in the fixed feature order shown in Table 9. Estimates and interval endpoints are serialized to four decimal places. The implementation assigns zero if a pooled standard deviation is zero. There is no significance-based feature selection or multiple-testing correction in this shared 20-anchor construction; intervals crossing zero are retained.

Table 9: The 20 stored statistical anchors. IDs are suffixes of `ML-COMMON-`; positive effects indicate higher mean values in the cancer group. Intervals are marginal bootstrap intervals, not simultaneous confidence bounds.

ID	CBC field	d_j	Lower 2.5%	Upper 97.5%
01	White blood cell count	0.3416	0.2046	0.4794
02	Neutrophil percentage	0.4293	0.3082	0.5477
03	Lymphocyte percentage	-0.5653	-0.7063	-0.4504
04	Monocyte percentage	0.2911	0.1617	0.4228
05	Eosinophil percentage	0.0187	-0.0826	0.1408
06	Basophil percentage	-0.1971	-0.2876	-0.0998
07	Neutrophil absolute count	0.3084	0.1625	0.5610
08	Lymphocyte absolute count	0.0785	-0.1550	0.2142
09	Monocyte absolute count	0.2624	0.1930	0.3429
10	Eosinophil absolute count	0.1033	0.0077	0.3230
11	Basophil absolute count	0.0110	-0.0882	0.1383
12	Red blood cell count	-1.3111	-1.4275	-1.1904
13	Hemoglobin	-1.1828	-1.2936	-1.0753
14	Hematocrit	-1.1692	-1.2905	-1.0630
15	Mean corpuscular volume	0.2991	0.1969	0.4132
16	Mean corpuscular hemoglobin	0.0646	-0.0488	0.1803
17	Mean corpuscular hemoglobin concentration	-0.4522	-0.5586	-0.3393
18	Red cell distribution width, CV	0.5854	0.4686	0.7020
19	Red cell distribution width, SD	0.7709	0.6320	0.9194
20	Platelet count	0.1755	0.0510	0.3233

These anchors summarize training-cohort associations. They are not learned diagnostic thresholds, calibrated patient-level probabilities, or causal effects. The same 20 statements are shared across the two anchor-enabled protocols. An anchor-disabled initialization instead uses the remaining 172 knowledge items and reconstructs its case-derived memory from that condition; it is not obtained by deleting anchors only at test time.

B Memory Representation and Operational Workflow

Return to appendix contents

B.1 Memory Card and Audit Schemas

The implementation separates restricted episode records (L0), reusable cards (L1), and an active-card catalog (L2). L0 preserves the learning input, reference, structured prediction, feedback, retrieval trace, and model-usage metadata. A candidate record and its subsequent decision share the originating event identifier. These records support reconstruction of the write history; the retriever never expands L0 episodes into future inference context. Table 10 summarizes the serialized L1 fields that connect reusable guidance to its provenance and retrieval state.

The schema rejects undeclared fields, bounds confidence to $[0, 1]$, and requires positive versions and token estimates. Candidate validation additionally requires nonempty semantic fields and retrieval

Table 10: Implemented L1 card fields grouped by their role in the method. Counters and confidence are operational metadata, not calibrated clinical reliability estimates.

Role	Serialized fields and interpretation
Identity and semantics	`card_id`, `section`, `title`, `reusable_rule`, `applicability`, and `limitations`.
Retrieval and scope	`keywords` and `supported_labels`; labels describe the stated scope and are distinct from reference labels attached to a new case.
Provenance	`raw_event_ids`, `source_types`, and `source_trust` link the card to its supporting events and source annotations.
Lifecycle	`stable`, `status`, `version`, `created_at`, and `updated_at`; status admits ACTIVE, INACTIVE, or QUARANTINED.
Operational metadata	`confidence`, `helpful`, `harmful`, `neutral`, and `estimated_tokens`.

keywords for an edit, and a target identifier for an update. Its character limits are 120 for the title, 700 for the rule, and 500 each for applicability and limitations; at most 12 keywords and eight supported labels are permitted. These structural checks establish a well-formed proposal, not its medical correctness.

The section vocabulary comprises legacy knowledge, public safety, statistical anchors, CBC patterns, non-tumor confounders, tumor ranking, correction rules, and safety boundaries. A section is a declared routing category; assignment is not an independently validated clinical classification. L2 groups ACTIVE cards by section and stores their identifiers, counts, aggregate token estimates, and up to 40 distinct keywords in lexicographic order. The catalog is reconstructed from card state, preventing an inactive card from remaining eligible through a stale catalog entry.

B.2 Update and Evaluation Procedures

For each designated learning case, the reasoner receives the case fields and packed L1 cards and returns a structured hierarchical prediction. The reference is supplied to the feedback and reflection components after prediction; it is absent from the reasoner's messages. Reflection receives the case, reference answer, prediction, feedback, and its packed cards, and proposes at most one ADD, UPDATE, or NO_CHANGE operation. It is instructed to abstract a transferable pattern, confounder, correction, ranking cue, or limitation rather than reproduce the patient panel. The configured forced-candidate variant requires ADD or UPDATE for each usable learning case, but does not bypass downstream validation.

Candidate generation and memory mutation are separate operations. The controller validates the proposal, records its provenance, computes an admission decision, applies permitted changes, and enforces active-memory capacity. Both admitted and withheld proposals retain candidate and decision records. The reasoner and reflector can retain different card subsets because their complete prompts are fitted separately. Held-out inference terminates after prediction and has no reflection or write operation.

B.3 Extended Feedback Decomposition

Let d_t and z_t denote the reference disease and tumor indicators. The implemented feedback uses reference-defined applicability,

$$\boldsymbol{\alpha}_t = (1, \mathbf{1}[d_t = 1], \mathbf{1}[z_t = 1]), \qquad F_t^{(s)} = \begin{cases} \mathrm{Eval}_s(y_t, g_t), & \alpha_t^{(s)} = 1, \\ \perp, & \alpha_t^{(s)} = 0. \end{cases} \tag{10}$$

At S1, the evaluator compares the predicted disease status with the reference status. At S2, it compares tumor-risk status only for reference disease cases. At S3, it records both equality of the first ranked group to the reference group and membership of the reference group in the three-group shortlist, only for reference tumor cases. Inapplicable outcomes are serialized as null. A valid abstention or an earlier routed stop supplies no downstream label, so an applicable comparison is false rather than omitted. The prediction contract requires three distinct permitted groups when tumor risk is asserted and no tumor ranking when the route stops earlier. Unparseable or schema-invalid responses

must pass the provider's structured-response handling before feedback can be computed; they are not silently converted into valid learning predictions. Missing or unknown reference categories fail canonical target construction.

The reflector also proposes helpful, harmful, or neutral tags for retrieved dynamic cards. The controller accepts at most one tag per eligible card and discards tags for stable, inactive, or non-retrieved cards. Accepted tags increment the corresponding counters even when no new rule is admitted. These are outcome-conditioned reflector judgments, not measured causal contributions of individual cards. They inform later capacity management and batch representative selection; they are not terms in the query-relevance score. Task feedback and these utility annotations do not instantiate the separate claim-faithfulness evaluation reserved in Appendix D.

B.4 Write Decisions and Active-Memory Semantics

Admission checks and precedence. Before cross-card comparison, the controller rejects proposals that copy a case identifier, contain a shared span of at least 14 normalized tokens from the case or reference answer, repeat multiple parsed patient values, exceed the configured card budget, or introduce standalone numeric rules from a single case. These are deterministic text checks, not a complete de-identification or factuality guarantee. Generic or very short rules also receive non-blocking quality flags.

For a structurally valid and copy-safe proposal, the ontology-aware decision order is as follows. NO_CHANGE terminates without a semantic edit. Internal inconsistency flags, including unknown supported labels, contradictory matched predicates, or specified text–label mismatches, produce HUMAN_REVIEW. A targeted UPDATE is then accepted only if the target is an active dynamic card; otherwise it produces REJECT_INVALID. For ADD, an identical normalized rule in stable knowledge produces REJECT_DUPLICATE, whereas an identical rule in dynamic memory produces MERGE_EXISTING. Exact matches are resolved in the current card-list order. Relaxed-build and Always modes then admit other valid additions directly.

The remaining gated ADD proposals are compared with active dynamic cards in the same section. Candidate pairs are ordered by exact normalized-title match, decreasing rule, title, and keyword Jaccard similarities, and finally card identifier. Write J_r, J_t, and J_k for these three similarities. A pair enters the scope check if $J_k \geq \tau_c$, $J_t \geq \tau_t$, or $J_r \geq \tau_r$, where thresholds are configuration parameters. Within each eligible pair, matched opposite predicates with sufficient topic-token overlap yield QUARANTINE_CONFLICT; unresolved mutually exclusive classification branches yield HUMAN_REVIEW when ontology review is enabled. Labels in the same tumor branch, or a disease ancestor and its descendant, are not deemed contradictory solely because their strings differ.

When per-case semantic deduplication is enabled, merging is triggered by an exact title with $J_r \geq \tau_e$, by $J_t \geq \tau_t$ together with $J_r \geq \tau_r$, by $J_r \geq \tau_{rr}$, or by $J_k \geq \tau_d$. The first decisive pair determines the action; unmatched proposals yield WRITE_NEW. Here "semantic" deduplication denotes deterministic comparisons of authored semantic fields, not embedding similarity or a learned entailment classifier. Predicate matching covers specified textual patterns and cannot establish the absence of arbitrary contradictions.

Application of admitted changes. WRITE_NEW creates an active dynamic card whose identifier is derived from a hash of the event, section, and rule. MERGE_EXISTING preserves the target's model-visible rule, applicability, and limitations while combining keywords, supported labels, event links, and source types, taking the larger confidence, and incrementing its version and neutral counter. The v3.1 adapter retains the existing token estimate when this visible text is unchanged. UPDATE_EXISTING replaces the semantic fields of the target, retains accumulated provenance, and increments its version. Stable cards are excluded from update targets.

The admission and withholding sets are

$$\mathcal{A}_{\text{admit}} = \{\text{WRITE_NEW}, \text{MERGE_EXISTING}, \text{UPDATE_EXISTING}\}, \tag{11}$$

$$\mathcal{A}_{\text{withhold}} = \{\text{NO_CHANGE}, \text{REJECT_INVALID}, \text{REJECT_DUPLICATE}, \text{QUARANTINE_CONFLICT}, \text{HUMAN_REVIEW}\}. \tag{12}$$

$$\mathcal{E}_{t+1}^{\text{semantic}} = \begin{cases} \text{Apply}(\mathcal{E}_t, \Delta_t), & a_t \in \mathcal{A}_{\text{admit}}, \\ \mathcal{E}_t, & a_t \in \mathcal{A}_{\text{withhold}}. \end{cases} \tag{13}$$

This equation describes semantic admission, separate from utility-counter updates and capacity maintenance. Withheld candidates remain in the audit records and are not instantiated as active cards. In particular, QUARANTINE_CONFLICT is an action label; the online application routine does not create a new QUARANTINED card or deactivate the existing conflicting card. HUMAN_REVIEW records a review requirement without implementing expert adjudication or automatic later admission.

Two further boundaries are material. A valid targeted UPDATE returns before the ADD similarity scan, and the runtime target check establishes active dynamic membership rather than independently enforcing the prompt's retrieved-target restriction. Exact-rule matching also precedes the broader conflict scan. Thus, the implemented checks should not be interpreted as an exhaustive consistency proof for every admitted edit.

B.5 Section Retrieval and Context Packing

Let Q be the set of normalized query tokens, W_s the section's stored keyword set, W_i a card's keywords, and T_i its normalized title tokens. Normalization case-folds text and extracts alphanumeric/underscore tokens and individual Chinese characters. The implemented scores are

$$R_s = |Q \cap W_s|, \qquad R_i = |Q \cap W_i| + |Q \cap T_i| + 0.2\,\mathbf{1}[i \in \mathcal{K}]. \tag{14}$$

Sections are sorted by decreasing R_s, breaking ties by catalog identifier. Active cards in the selected sections are sorted by decreasing R_i, breaking ties by card identifier. Zero-overlap sections can remain eligible when needed to fill the configured section count; there is no learned section selector or L0 fallback.

In the separate-cap retrieval adapter, the ranked list is divided into stable and dynamic candidates. Stable candidates are capped at k_K; dynamic candidates must additionally overlap the query in a keyword or title token and are capped at k_E. The adapter considers the stable list followed by the dynamic list, greedily admitting complete cards that fit the memory budget and skipping oversized cards. The resulting order therefore reflects both within-type relevance and explicit stable-first packing. No minimum dynamic-card quota is enforced by this adapter. Lexical overlap is an eligibility proxy, not verification that a card's full applicability conditions hold for the patient.

$$\widehat{\mathrm{Tokens}}(\mathrm{Prompt}(x_t, \mathcal{P}_t)) \le b, \qquad |\mathcal{P}_t \cap \mathcal{K}| \le k_K, \qquad |\mathcal{P}_t \cap \mathcal{E}_t^{\mathrm{active}}| \le k_E. \tag{15}$$

Card storage uses the approximation $\max(1, \lceil \mathrm{UTF8Bytes}(\text{text})/3 \rceil)$. A second, provider-level estimate includes the complete messages and output-schema instructions. If that estimate exceeds the hard input limit, the controller removes whole cards from the tail and rebuilds the prompt until it fits. Case fields and retained rules are not truncated. If the prompt cannot fit with no cards, provider limit handling stops the request rather than silently dropping patient information. The estimate is not asserted to equal the backbone tokenizer's exact count; the inference provider also checks returned input-token usage and reserves output capacity in its context configuration. Selected sections, card identifiers, estimated usage, and both budget-driven and hard-limit omissions are recorded for attribution.

B.6 Recursive Evolution and Read-Only Evaluation

The case stream changes external memory rather than the frozen model parameters. The implemented transition can be decomposed into utility updates, semantic admission, capacity control, and an optional declared batch-consolidation boundary. If active dynamic tokens exceed the configured capacity, cards are demoted to INACTIVE in ascending order of helpful-minus-harmful count, confidence, update timestamp, and identifier until the budget is met. Stable cards are outside this eviction pool; inactive cards retain their provenance.

Batch consolidation groups active dynamic cards by section and identical normalized title. Beyond the contradiction and label-branch checks below, this batch path does not impose an additional rule-similarity threshold. A representative is chosen by decreasing helpful-minus-harmful count, helpful count, total utility count, and confidence, followed by increasing creation time and identifier. Compatible duplicates transfer their counters and provenance to that representative and become inactive. Explicit predicate conflicts or unresolved exclusive-branch ambiguity instead generate a

batch review record and leave the existing cards active. This behavior differs from withholding a newly proposed card at admission: batch review does not retroactively quarantine all implicated memory.

For a sequential update stream, these operations define U_t and

$$\mathcal{M}_T = (\mathcal{U}_{T-1} \circ \cdots \circ \mathcal{U}_0)(\mathcal{M}_0), \qquad \hat{y}_i^{(T)} = f_\theta(x_i, \mathrm{Context}(x_i, \mathcal{M}_T)) \,. \tag{16}$$

Snapshots serialize the cards and regenerated catalog, with separate SHA-256 hashes of their canonical representations. The experiment implementation additionally binds the producing configuration, parent snapshot, anchor condition, and completed-case count, and verifies a sealed parent before use. Its prediction branch loads that parent, performs retrieval and structured reasoning without learning targets, and commits predictions rather than modified cards; a subsequent parent-seal check detects changes to the bound memory files. This separation prevents held-out predictions from entering the update recurrence. It does not, by itself, resolve how investigators select configurations or checkpoints from reported results.

C Implementation Details and Prompt Templates

Return to appendix contents

C.1 Backbone, Context, and Decoding Configuration

The principal backbone is Qwen3.8-27B, served through the Ollama chat interface with frozen weights. The 16K configuration specifies the Q4_K_M model variant, temperature zero, and disabled thinking output. The model artifact is identified by a digest rather than its display name alone. Reasoning and reflection use the same configured backbone, while their roles and output contracts are supplied through separate messages. No gradient update is part of memory adaptation. Table 11 gives the stable- and dynamic-card limits for the two context profiles selected for this experiment. These manually specified comparison settings do not restrict the method to two window sizes: the context limit is configurable within the backbone and inference service's supported capacity, while retrieval and complete-prompt fitting automatically adapt the supplied memory to each case within the configured budget.

Table 11: Card-count limits for the two manually selected experimental context profiles. These settings are comparison configurations, not the only supported capacities. The limits are upper bounds; retrieval and full-prompt fitting automatically determine the realized card counts within the configured budget.

Context profile	Maximum stable cards	Maximum dynamic cards
4K	8	4
16K	12	4

For the 16K memory-evolution configuration, the requested context is 16,384 tokens, the hard input limit is 15,360, and the maximum generated output is 768 tokens. Their sum leaves 256 tokens of headroom. Retrieval selects at most four sections and budgets 2,400 estimated tokens for memory; individual candidate cards are limited to 240 estimated tokens, and active dynamic memory to 12,000. The latter is a storage bound and does not imply that all active memory is placed in a prompt. Stable and dynamic cards share the retrieval token budget; instructions, the case, serialization, and schema overhead are accounted for by full-prompt fitting, without a fixed per-field allocation.

The inference adapter requests schema-constrained JSON, validates the returned object and hierarchical route, and permits at most one structural repair in this configuration. For an invalid three-group ranking, the repair adapter can expose three separate rank fields, then map them back to the canonical array and validate the result. This changes the response representation, not the target label or the patient's evidence. The configuration permits one retry for transient transport failures. Truncated or over-budget outputs are not accepted as completed structured predictions. Temperature zero and a fixed seed specify decoding settings but do not establish bitwise repeatability across inference engines or hardware. Top-p is not explicitly overridden by the inspected configuration and is therefore not assigned an invented numerical value here.

C.2 Baseline Adaptations and Matched Budgets

The baseline adaptations preserve the same disease, tumor-risk, and three-group ranking task while changing the source and evolution of guidance. Frozen_LLM uses a fixed medical instruction with empty learned guidance. Static RAG retrieves from a fixed knowledge-card collection and performs no feedback-driven revision; its historical implementation uses the section/card retrieval and prediction contract of the memory framework. The context-optimization baselines serialize their guidance alongside the current patient fields and return a shared four-field prediction object. This contract omits the confidence and copied-observation fields of the proposed method while retaining its classification and ranking decisions.

Dynamic Cheatsheet. DC maintains a single cumulative text. Each learning case is first answered under the current cheatsheet, after which a curator proposes a complete replacement using the previous text and the model answer. The adaptation explicitly treats model answers as unverified and does not supply canonical case-reference feedback to this curator. Approved initialization knowledge can be incorporated through the same bounded-text interface. A replacement exceeding the memory budget or failing its output contract is rejected, preserving the previous text. Thus, DC adapts by whole-text revision rather than provenance-linked local card admission.

ACE. ACE represents guidance as a collection of identified skills. A reflector receives the current skillbook and approved knowledge or canonical learning feedback, emits error identification, root-cause analysis, a corrected approach, an insight, and utility tags. A separate SkillManager proposes ADD, UPDATE, and REMOVE operations. The controller assigns new identifiers, validates referenced identifiers and operation fields, rejects exact duplicates, and bounds the complete resulting skillbook including metadata. Valid utility tags increment skill counters. Operations are applied to a copy and committed only if the resulting structure is valid and within capacity. These structural and capacity checks do not introduce the proposed method's clinical conflict Gate.

GEPA. GEPA optimizes a reusable medical-guidance component using the vendored Pareto-search implementation. Initialization knowledge supplies the initial instruction. Initialization cases are partitioned by patient group and stratified by canonical label into internal training and validation subsets, targeting a validation fraction of 0.2; singleton strata remain in training. The update phase adds update cases to the training subset while retaining the internal validation set. The reward averages correctness over applicable hierarchical stages, with S3 using the first ranked group and upstream routing misses receiving zero. Reflection proposes bounded instruction revisions from training traces and reference feedback. Search uses an instance-level Pareto frontier, reflection mini-batches of three, and configured metric-call budgets of 5,000 for initialization and 10,000 for updating; merge proposals are disabled. Held-out test cases do not participate in this internal optimization.

The DC, ACE, GEPA, and Frozen_LLM configurations use a 16,384-token context, a 4,096-token output ceiling, and a 512-token prompt safety margin, with temperature zero, and thinking disabled. DC, ACE, and GEPA bound guidance at 2,400 estimated tokens. Their shared inference wrapper allows one structure-only repair preserving existing route decisions; an invalid prediction after this attempt becomes an explicit abstention. Equal task contracts or context-window sizes do not imply equal inference costs: DC adds one curator call per learning case, ACE adds reflector and SkillManager calls, and GEPA evaluates multiple candidate instructions. Their output reservation also differs from the proposed method's 768-token configuration. Consequently, the implementations support a comparison of adaptation mechanisms, but a claim of identical computational budgets would require additional matching beyond the settings stated here.

C.3 Exact Prompt Documentation

The following panels reproduce the operative instruction text and message boundaries, with line wrapping adjusted for readability. Braced placeholders denote substituted data, not additional instructions. The proposed method serializes each retrieved card as its identifier and section followed by title, rule, applicability, and limitations. An empty case context is rendered as `(empty)`. A provider-level system message requests one JSON object matching the generated response schema; schema constraints and route validation complement the natural-language instructions. Formatting and placeholder notation below do not change these message contents.

Reasoner. The reasoner sees no learning reference. The panel shows the inference/update abstention policy. During initialization, the policy sentence is replaced by `This is memory construction: return ANSWER for every syntactically usable input.` If no card is retrieved, the memory message contains `No active memory card was retrieved.`

Prompt C.1: Reasoner prompt and message layout.

```
System: task
Closed-set CBC benchmark, not clinical advice. Follow three stages: (1)
NO_DISEASE vs DISEASE; (2) for DISEASE, NON_TUMOR_DISEASE vs TUMOR_RISK; (3) for
TUMOR_RISK, rank exactly three distinct groups from C16, C18/C20, C22, C34,
C53, C56, C61, C91/C92. ABSTAIN or HUMAN_REVIEW is allowed for malformed,
insufficient, or unresolved input. Copy at most three analyte/value
observations exactly from CASE_INPUT. Emit no free-text rationale.
System: memory
ACTIVE_MEMORY_CARDS:
[{card_id}|{section}] {title}. Rule: {reusable_rule} Applies when:
{applicability} Limits: {limitations}
{additional_card_lines}
User
CASE_INPUT:
{question}
CONTEXT:
{context}
```

Reflector. The reflector receives reference information only for a designated learning case. Empty retrieved-card and eligible-identifier lists are rendered as `(none)`. Its memory view is packed independently from the reasoner's view. The instruction panel uses the optional-edit policy; the configured forced-candidate variant substitutes the sentence shown in the subsequent panel. The word-length requests in the instruction coexist with the character bounds enforced by the candidate schema in Appendix B.

Prompt C.2: Reflector system instruction.

```
System
You are the Reflector and SkillManager in a persistent-memory experiment. The
complete case and reference answer below are restricted reflection inputs, not
text to copy. Propose at most one reusable memory edit. Generalize a
reasoning pattern, correction, confounder, ranking cue, or safety boundary.
Do not reproduce the patient panel, age, complete reference answer, or a
patient-specific conclusion. Do not invent thresholds or causal/diagnostic
rules from one case. Independently tag each retrieved dynamic experience card
that was actually used as helpful, harmful, or neutral for this outcome; never
tag an ID outside DYNAMIC_CARD_IDS_ELIGIBLE below. If that list is (none),
memory_utility_tags must be empty. Utility tags do not require a new edit. If
no transferable learning exists, return NO_CHANGE. Even for NO_CHANGE, still
return any evidence-based utility tags. Keep the title within 12 words,
reusable rule within 60 words, applicability within 35 words, and limitations
within 35 words. UPDATE may name only an existing retrieved card.
BUILD_RELAXED affects the later deterministic Gate and does not permit verbatim
copying.
```

Prompt C.3: Reflector data messages and conditional policy.

```
System: retrieved context
RETRIEVED_CARDS:
[{card_id}] {title}. Rule: {reusable_rule} Applies when: {applicability}
Limits: {limitations}
{additional_card_lines}
System: eligible tag targets
DYNAMIC_CARD_IDS_ELIGIBLE_FOR_UTILITY_TAGGING: {comma_separated_dynamic_card_ids}
User
GATE_MODE={gate_mode}
CASE_INPUT={question}
CONTEXT={context}
PREDICTION={prediction_json}
STRUCTURED_FEEDBACK={feedback_json}
REFERENCE_ANSWER_REFLECTION_ONLY={reference_answer}
Forced-candidate policy replacement
This is the provisional all-experience development arm. For every usable
case, return exactly one ADD or valid UPDATE; NO_CHANGE is not permitted. The
experience may be a case-derived correction or caution and must state its
single-case limits.
```

Structured response contracts. The panels specify field types rather than fabricate example patient outputs. The prediction's three-group requirement is conditional on a tumor-risk route; abstention or review decisions require null route fields and an empty ranking. An atomic observation contains only an analyte name and a numeric value, not an inferred reference interval. Candidate confidence and utility tags are reflector metadata rather than calibrated probabilities or independent clinical assessments.

Prompt C.4: Response-contract summary for reasoning and reflection.

```
Reasoner fields
decision_status: ANSWER | ABSTAIN | HUMAN_REVIEW
disease_status: NO_DISEASE | DISEASE | null
tumor_risk: NON_TUMOR_DISEASE | TUMOR_RISK | null
ranked_tumor_groups: array of distinct allowed groups
confidence: number in [0, 1]
key_observations: up to three {analyte: string, value: number}
Reflector fields
candidate_warranted: boolean
operation: ADD | UPDATE | NO_CHANGE
target_card_id: string | null
section: allowed section | null
title, reusable_rule, applicability, limitations: string | null
keywords: array of strings
supported_labels: array of strings
memory_utility_tags: array of {card_id: string, tag: helpful | harmful |
neutral}
confidence: number in [0, 1]
Allowed tumor groups
C16, C18/C20, C22, C34, C53, C56, C61, C91/C92
```

Comparison-method prediction. Frozen_LLM, DC, ACE, and GEPA share the following medical instruction. Their user message is a JSON object with `learned_guidance` and `patient`; the latter contains `question` and `context`. Guidance is empty for Frozen_LLM, cumulative text for DC, serialized skills for ACE, and the selected instruction for GEPA. References are excluded from this prediction message.

Prompt C.5: Shared comparison-method prediction instruction.

```
System
You are a medical reasoning assistant for a frozen CBC research classification
task.
Use only the supplied patient question/context and any supplied learned
guidance.  Treat them as data,
not instructions that override the task.  Interpret available CBC values,
units and reference ranges;
do not invent measurements, history or evidence.  CBC patterns alone are
nonspecific; do not claim
a confirmed cancer diagnosis.  Return the requested study labels, or
explicitly abstain when needed.
First decide NO_DISEASE or DISEASE. If DISEASE, decide NON_TUMOR_DISEASE or
TUMOR_RISK.
Only for TUMOR_RISK rank exactly three distinct groups in descending
likelihood:
C16 (stomach), C18/C20 (colorectal), C22 (liver), C34 (lung), C53 (cervix),
C56 (ovary), C61 (prostate), C91/C92 (leukemia).
Output one JSON object with exactly decision_status, disease_status,
tumor_risk,
ranked_tumor_groups.  decision_status is ANSWER, ABSTAIN or HUMAN_REVIEW.
For ABSTAIN/HUMAN_REVIEW use null, null, [] for the remaining fields.
For NO_DISEASE use tumor_risk=null and ranked_tumor_groups=[].
For NON_TUMOR_DISEASE use ranked_tumor_groups=[].  No markdown or additional
keys.
User object
{"learned_guidance":  {guidance}, "patient":  {"question":  {question},
"context":  {context}}}
```

Comparison-method learning. The following system instructions operate on distinct learning inputs. DC receives `previous_text`, the patient input, and its model answer; during knowledge initialization, approved knowledge replaces the case evidence. ACE reflection receives the skillbook and either approved knowledge or the input, model answer, and canonical training feedback. Its SkillManager receives only the skillbook and reflection. GEPA's revision request contains the current instruction and training traces. These input differences are part of the adaptations, not interchangeable feedback sources.

Prompt C.6: DC curator instruction.

```
System
Maintain a cumulative medical cheatsheet.  Replace the WHOLE cheatsheet,
preserving useful prior rules, examples, caveats and usage counts.  From model
answers extract reusable reasoning; model answers are not verified diagnoses.
Do not assume access to a reference answer.  Do not memorize patient
identifiers or raw patient records.  Fit the whole text within approximately
2400 estimated tokens; be concise.  Return only {"text":"complete new text"}.
```

Prompt C.7: ACE reflector instruction.

```
System
You are ACE Reflector.  Analyze errors/success and extract a reusable, bounded
medical learning.  Use only approved knowledge or canonical training feedback,
never infer a population-level medical rule from a single case.  Tag only
relevant existing skills by id as helpful/harmful/neutral.  Return the exact
requested JSON fields.
```

Prompt C.8: ACE SkillManager instruction.

```
System
You are ACE SkillManager. Apply incremental changes to the skillbook. Return
operations ADD/UPDATE/REMOVE; ADD id is empty, UPDATE/REMOVE id must exist.
All operations need string id, section, content (REMOVE content/section can be
empty). Use an empty list when no new useful knowledge. Merge by UPDATE plus
REMOVE when needed. Preserve useful entries, avoid duplicate or
patient-specific memorization. The complete resulting skillbook must fit 2400
estimated tokens including metadata. This is a capacity limit, not a medical
conflict gate.
```

Prompt C.9: GEPA instruction-revision prompt.

```
System
Improve the reusable medical guidance based on execution traces and canonical
training feedback. Preserve the fixed disease/risk/three-group task contract.
Do not memorize individual cases. Keep text below 2400 estimated tokens.
Return only a JSON object with text.
```

The DC and GEPA text-update contract is a single `text` string. ACE reflection requires `error_identification`, `root_cause_analysis`, `correct_approach`, `key_insight`, and `skill_tags`. ACE management returns an `operations` array, each element containing `type`, `id`, `section`, and `content`. The numerical capacity in these panels is instantiated from the 2,400-token configuration; it is an estimated full-guidance bound, not a per-operation allowance. GEPA initializes guidance using a separate knowledge-synthesis instruction before candidate search.

Prompt C.10: GEPA knowledge-initialization instruction.

```
System
Synthesize a reusable medical task instruction from approved training
knowledge. Preserve useful prior guidance. This initializes the prompt for
subsequent GEPA optimization. Do not memorize patient identifiers or raw
patient records. Fit the whole text within approximately 2400 estimated
tokens; be concise. Return only {"text":"complete new text"}.
```

Deterministic components and review boundary. Stage feedback is computed by label comparison. Section/card retrieval, candidate checks, Gate decisions, and card application use the rules in Appendix B; they do not invoke separate model-based selector, audit, or archive agents. Accordingly, there is no additional prompt for these components. HUMAN_REVIEW is a recorded disposition, not evidence that an expert reviewed and readmitted a candidate. No reviewer qualifications, adjudication protocol, or completed review outcomes are inferred from that action name.

D Evaluation Metrics and Statistical Reporting

Return to appendix contents

D.1 Reference-Defined Stage Eligibility

Evaluation is a deterministic transformation of completed predictions and separately stored reference labels. Records are aligned by case identifier, rather than by an assumed row correspondence. The scoring interface rejects duplicate identifiers or unequal prediction/reference identifier sets. Reference records use one of three canonical risk categories: normal without disease, non-tumor disease, or tumor; only tumor references carry a secondary label from the eight-group vocabulary. Dataset registration and prediction-state bindings associate the scores with the intended test population and label version. Scoring does not invoke the language model or modify its memory snapshot.

Let $\mathcal{D}_s$ denote the reference-eligible population for stage s, and let d_i, z_i be its disease and tumor indicators. The populations are

$$\mathcal{D}_1 = \mathcal{D}, \qquad \mathcal{D}_2 = \{i \in \mathcal{D} : d_i = 1\}, \qquad \mathcal{D}_3 = \{i \in \mathcal{D} : z_i = 1\}, \qquad N_s = |\mathcal{D}_s|. \tag{17}$$

Thus, the MSA denominators are 780, 600, and 510, and the HMT denominators are 1,200, 800, and 400 (Appendix A). These denominators do not depend on which cases the model elects to answer or routes to a later stage.

The prediction contract permits ANSWER, ABSTAIN, and HUMAN_REVIEW. An ANSWER must specify a disease decision; a disease-positive route must additionally specify tumor risk. A tumor-risk-positive route requires exactly three distinct permitted tumor groups, in ranked order. Earlier stopping routes must not supply later-stage labels, and abstention or review decisions require null route fields and an empty ranking. Duplicate candidates, unknown group names, short lists, and inconsistent routes are invalid scored objects. The scorer neither pads nor reorders a shortlist and does not infer missing groups from free text. Response repair belongs to the inference interface described in Appendix C; a saved valid abstention is distinct from a malformed object that fails the scoring contract.

For a validated prediction, let a_{is} indicate that an output reaches stage s. S1 requires ANSWER, S2 additionally requires a predicted disease-positive route, and S3 additionally requires a predicted tumor-risk-positive route. For $i \in \mathcal{D}_s$, set $\hat{y}_{is} = \perp$ when $a_{is} = 0$. An upstream routing miss remains in the denominator for full-population metrics but is excluded from the denominator of conditional accuracy, as defined in Appendix D.2.

D.2 Accuracy, F1, Balanced Accuracy, and Coverage

For reference label y_{is} and fixed class set $\mathcal{C}_s$, the full-population metrics are

$$\mathrm{Acc}_s = \frac{1}{N_s} \sum_{i \in \mathcal{D}_s} \mathbf{1}[\hat{y}_{is} = y_{is}], \qquad \mathrm{Cov}_s = \frac{1}{N_s} \sum_{i \in \mathcal{D}_s} a_{is}, \tag{18}$$

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \qquad \mathrm{Bal.Acc}_s = \frac{1}{|\mathcal{C}_s|} \sum_{c \in \mathcal{C}_s} \frac{TP_c}{n_c}, \tag{19}$$

$$\text{Ma-F1}_s = \frac{1}{|\mathcal{C}_s|} \sum_{c \in \mathcal{C}_s} F1_c, \qquad \text{Wt-F1}_s = \sum_{c \in \mathcal{C}_s} \frac{n_c}{N_s} F1_c. \tag{20}$$

Here n_c is reference support, and counts are formed on the entire reference-eligible population. The class sets contain two disease labels at S1, two tumor-risk labels at S2, and the eight tumor groups at S3. A missing output contributes a false negative to its reference class but is not an additional clinical class in the macro averages. Consequently, balanced accuracy averages per-class recall, whereas weighted F1 weights each class by its reference support. F1 is computed within a run from its class counts, not reconstructed by averaging rounded accuracy or precision values.

For the ordered tumor shortlist $\mathrm{Top3}_i$, the scorer additionally computes

$$\mathrm{Acc@3} = \frac{1}{N_3} \sum_{i \in \mathcal{D}_3} \mathbf{1}[y_{i3} \in \mathrm{Top3}_i]. \tag{21}$$

S3 Acc and Acc@1 are the same first-ranked-group quantity. Acc@3 credits a case at most once and remains zero for a route miss. Ranking position matters for Acc@1 but not for membership in Acc@3; neither is a graded ranking metric.

Let $A_s = \sum_i a_{is}$ and $K_s = \sum_i \mathbf{1}[\hat{y}_{is} = y_{is}]$. The implementation retains the conditional diagnostic $\mathrm{Acc}_s^{\mathrm{routed}} = K_s/A_s$, and its S3 counterpart divides top-three hits by the number of emitted tumor rankings. For $A_s > 0$,

$$\mathrm{Acc}_s = \mathrm{Cov}_s \, \mathrm{Acc}_s^{\mathrm{routed}}. \tag{22}$$

Conditional accuracy describes cases with valid stage-specific outputs and is distinguished from full-population accuracy in the result tables. It should be interpreted jointly with response coverage. Coverage quantifies output availability, not correctness, clinical calibration, or evidence faithfulness.

Undefined class precision, recall, and F1 are assigned zero within a nonempty stage population. Fixed-class macro averages retain those zeros when a class has no support, rather than silently reducing the class vocabulary. An entirely empty stage population yields null aggregate rates and F1 values in the current scoring wrappers; a stage with no answers also has null routed accuracy. This distinguishes unavailable metrics from a measured zero. The shared hierarchical implementation

retains rates as fractions, rounds stored class/rate values to twelve decimal places, and converts them to percentages only for presentation. Display precision is separate from the underlying counts.

The proposed-method scoring wrapper independently recomputes the same metrics from integer counts and requires agreement with the hierarchical scorer to within 2×10^{-11} for defined rates. It records total, answered, correct, top-three-correct, and per-class counts alongside the scores. The baseline scorer preserves the same routing, fixed-class averaging, and empty-population conventions, despite its smaller four-field prediction schema. This provides a common metric definition without requiring identical auxiliary model-output fields.

D.3 Checkpoint Aggregation and Paired Comparisons

A score record binds a prediction set to its protocol, anchor condition, method, checkpoint, model/retrieval settings, reference-label hash, and metric contract. The within-design collector additionally checks population identity, compatible initialization and update profiles, and within-method batch schedules. Missing or duplicate declared cells are not silently replaced by partial collections. A checkpoint comparison therefore concerns a specified memory state and configuration, not merely a displayed method name.

A selected checkpoint or observed condition is a descriptive selection from an identified score record. Each reported row retains the protocol, condition, checkpoint, and configuration from which its metrics were obtained. Selected-result reporting is therefore descriptive and should not be interpreted as an inferential estimate after model or checkpoint selection.

For two conditions evaluated on the same reference population, a descriptive difference is $\Delta_s = 100(m_{s,B} - m_{s,A})$ percentage points when rates are stored as fractions. Population matching supports an aligned comparison but does not itself supply a confidence interval or a significance test. The current cross-model collector checks the reference population and selected memory identity and retains configuration metadata; it does not average MSA and HMT into one population. The implemented collectors do not produce paired bootstrap tests, multiplicity-adjusted significance levels, or causal estimates of memory benefit. Such inferential claims require a separately specified sampling unit and analysis.

D.4 Claim-Level Faithfulness: Reserved Evaluation

The task scorer and the auxiliary evidence-consistency audit answer different questions. The former compares classification outputs with reference diagnoses. The latter operates on explicitly structured claims and their corresponding case inputs; it does not extract all factual propositions from unrestricted natural-language reasoning. In the implemented audit, a numeric `key_observation` is supported if the named analyte has an available input value and the cited number matches it with absolute tolerance 10^{-9} and zero relative tolerance. An absent analyte/value is unsupported, while a differing available numeric value is contradicted.

The audit also supports two legacy output surfaces. A laboratory-direction statement with an available analyte is unverifiable when no valid patient-specific reference interval is supplied; if the analyte is absent, it is unsupported. A nonempty unstructured safety note is recorded as an unverifiable text surface, without claiming a complete atomic decomposition. These fields are not required by the current reasoner contract. In particular, the four-field baseline predictions do not emit numeric observations, so the audit cannot establish a matched claim-faithfulness comparison by treating absent claims as correct ones.

Let S, C, U, V count supported, contradicted, unsupported, and unverifiable audit records. The implemented micro-aggregated rates are

$$P_{\text{claim}} = \frac{S}{S + C + U}, \qquad E_{\text{claim}} = \frac{C + U}{S + C + U}, \tag{23}$$

$$E_{\text{strict}} = \frac{C + U + V}{S + C + U + V}, \qquad R_{\text{unverifiable}} = \frac{V}{S + C + U + V}. \tag{24}$$

These expressions aggregate emitted records across cases, rather than averaging case-level ratios. Zero denominators yield null, not a perfect precision or zero error claim. Cases with no emitted records add no claim counts and remain in the denominator of claim-emission case coverage. The

audit also reports coverage of available input values, counting unique supported case–analyte pairs relative to all parsed nonmissing input fields. This coverage is an anti-silence diagnostic, not recall of clinically necessary findings. Repeated emitted records contribute to the status totals, whereas input-value coverage deduplicates supported case–analyte pairs.

The strict rate counts unverifiable surfaces in its numerator as a sensitivity convention; it does not establish that each such surface is medically false. The numerical-transcription diagnostic counts contradicted numeric observations relative to emitted numeric observations, while the broader claim-error rate also includes unsupported observations. Claim recall and claim F1 remain null without an annotated required-claim set. Expert-adjudicated medical hallucination is likewise not supplied by numeric input matching. The present reliability measures therefore remain balanced accuracy and response coverage; no additional claim-level result is inferred here.

D.5 Conditional Count-Consistency Checks

Integer counts provide a common representation for accuracy, coverage, and binary top-three hits. For a nonempty stage, let H_s be top-three hits where applicable, and let $M_{c,j}$ count reference class c with predicted class j, including a separate missing-output column $j = \perp$. Then

$$\sum_j M_{c,j} = n_c, \qquad \sum_c n_c = N_s, \qquad A_s = N_s - \sum_c M_{c,\perp}, \qquad K_s = \sum_c M_{c,c}. \tag{25}$$

These identities connect the scorer's per-class and missing-output counts to a row-normalized confusion-matrix presentation. A Miss column denotes an unavailable stage output caused by abstention, review, or upstream stopping; it is not an additional reference diagnosis. Displayed row percentages should use the same reference support as the scalar metrics. They need not sum to exactly 100 after independent decimal rounding.

For an individual score record, binary stage metrics are computed from integer counts over the fixed reference-eligible population. In particular, stage accuracy, response coverage, and top-three accuracy (where applicable) are determined by the corresponding correct, answered, and top-three-hit counts together using N_s for full-population rates and A_s for conditional accuracy. Count-level consistency is verified on the originating score record before any checkpoint summary is formed.

Displayed summary values should therefore be interpreted according to the statistic associated with their bound score records and the aggregation procedure specified in Appendix D.3, rather than reverse-engineered from rounded percentages alone. The manuscript retains the source denominator, summary rule, and prediction/reference bindings needed to distinguish a single-run proportion from a descriptively selected record.

Consistency is established by the originating score record, its denominator and summary rule, and its prediction/reference bindings. Figure cells, metric tables, and longitudinal curves can then be derived from the same numerical representation. Numerical agreement after rounding is a useful consistency condition, but cannot identify an unknown run or repair a missing provenance link. This section defines that relationship without changing the supplied manuscript results.

E Supplementary Results and Experimental Details

Return to appendix contents

E.1 Complete Endpoint and Longitudinal Results

Table 12 summarizes the final Ours values reported in Table 1. These task-level results complement the longitudinal view in Fig. 2: the summary compares methods on the reported metrics, whereas the trajectories describe performance as the available memory changes. S3 accuracy denotes the first-ranked tumor group, and Acc@3 denotes inclusion in the three-group shortlist. The stage populations and treatment of missing outputs follow Appendix D.

The checkpoint mechanism separates memory construction from subsequent updating. The zero-update state already contains the initialized stable knowledge and the experience acquired from initialization cases. Later checkpoints record the state after 100, 300, 500, or 700 additional update

Table 12: Summary of the reported Ours task scores (%), reproduced from Table 1.

Protocol	S1 Acc	S2 Acc	S3 Acc	S3 Acc@3
MSA	87.43	70.53	22.63	43.55
HMT	82.92	50.06	21.20	44.41

cases. Each checkpoint stores the cards, regenerated catalog, completed-case count, and parent-state bindings. Predictions load a fixed checkpoint and do not feed held-out answers back into memory.

Table 13 provides the numerical values underlying the checkpoint trajectories in Fig. 2. Each displayed value is taken from the descriptively selected score record bound to that protocol, checkpoint, and configuration; The same fixed reference-defined stage population is used at every checkpoint.

Table 13: Numerical supplement to Fig. 2: task scores (%) from the selected score records at the displayed memory checkpoints. Zero denotes the initialized memory before additional case updates. Four decimal places retain the stored plotting values

Protocol	Update cases	S1 Acc	S2 Acc	S3 Acc	S3 Acc@3
MSA	0	85.4704	68.9477	22.3529	42.2120
MSA	100	85.5868	69.6667	22.3880	42.7451
MSA	300	85.8395	69.9970	22.4290	42.9600
MSA	500	86.9231	70.1456	22.5280	43.1010
MSA	700	87.4277	70.5260	22.6256	43.5512
HMT	0	78.5000	42.6250	18.5000	41.0000
HMT	100	80.4073	44.2893	19.7088	41.8696
HMT	300	80.5737	44.5533	20.0260	42.2768
HMT	500	81.2380	44.8574	20.2146	42.6853
HMT	700	82.9243	50.0553	21.1984	44.4134

For a protocol, the same held-out cohort and reference-defined stage populations are retained across checkpoints. Consequently, changes along a trajectory concern the memory available to the reasoner rather than a changing test population. S3 Acc and Acc@3 remain distinct series because improved shortlist inclusion need not imply improved first-choice ranking. Stable knowledge remains read-only, while dynamic rules, their provenance, and their retrieval eligibility can change through the operations in Appendix B. The checkpoint representation captures their combined effect; it does not assume that every admitted update improves every subsequent prediction.

E.2 Matched Component Comparisons

Table 3 compares configurations along three implementation dimensions: statistical anchors, memory admission, and retrieval-context capacity. Their operational meanings are as follows.

Statistical anchors. The anchor-enabled condition includes the 20 training-derived statements described in Appendix A.3, together with the remaining 172 knowledge items. The anchor-disabled condition constructs its initial memory from those 172 items and performs initialization-case learning under that condition. The resulting dynamic memory is therefore conditioned on the knowledge available during learning. Anchor removal is not implemented as deleting 20 statements only from the final test prompt. Parent-snapshot checks retain the anchor condition when continuing an update stream or loading it for prediction.

Memory admission. The Gate condition applies the deterministic admission rules and its configured batch-consolidation schedule. The Always condition retains structural validation, copying restrictions, valid targeted updates, and exact-rule consolidation, while omitting the gated cross-card scan; its update configuration has no batch-consolidation schedule. Both conditions use frozen model parameters and preserve stable knowledge. The comparison therefore concerns these implemented memory-update policies, including their associated consolidation behavior. It does not

isolate an individual predicate or establish that all checks can be removed without changing the remaining workflow.

Context and retrieval capacity. The 4K and 16K profiles are two optional context settings manually selected for the experimental comparison, rather than intrinsic limits of the method. The context window and retrieval budget are configurable subject to the capacity supported by the backbone and inference service. During inference, the amount and composition of the supplied memory adapt automatically to the case through relevance ranking, budget-aware card selection, and complete-prompt fitting. Cards that exceed the available budget are omitted as whole units, preserving the case fields and retained rules. For the reported comparison, the window settings and card-count limits are specified in advance: the stable-card cap changes from eight to twelve, while the dynamic-card cap remains four. These controlled settings allow the experimental effects of context and retrieval allocation to be compared; the realized context can vary across cases within each setting.

These dimensions explain the adjacent comparisons in Table 3: adding anchors with Gate enabled, enabling the gated policy without anchors, and increasing context capacity without either component. The displayed configurations form a partial comparison rather than a full factorial design. They support interpretation of the reported configuration differences, without estimating all component interactions or attributing the entire difference to one low-level operation.

E.3 Cross-Backbone Transfer Protocol

Cross-backbone reuse transfers external memory rather than model weights. The source resource consists of Qwen-derived cards, their section catalog, and a manifest binding them to the source condition and checkpoint. Before recipient inference, the resource loader checks both file integrity and the logical hashes of the card and catalog representations. It also checks the corresponding protocol registry and anchor condition, so replacing the recipient model does not silently replace the underlying memory resource or test population.

The recipient uses the same case fields, card interface, and hierarchical task contract. Section/card retrieval is performed against the fixed source memory, and the retained complete cards are serialized into the recipient's prompt. The recipient's model identity and inference settings are recorded separately from the source-memory identity. Full-prompt fitting is performed for the recipient, so a shared source memory need not produce an identical final token sequence across different backbones. The retrieval record preserves selected and omitted card identifiers and the resulting context composition. Recipient prediction has no reflection, admission, or memory-reconstruction stage; the source resource remains unchanged throughout inference.

Table 2 reports Gemma 4 26B A4B (Google DeepMind, 2026) and MedGemma-27B (Sellergren et al., 2025) as recipients on MSA and HMT, with Qwen as the source-model reference. The shared representation makes memory consumable across backbones while leaving interpretation of that memory to each frozen model. Accordingly, the comparison describes task performance in the memory-reuse setting. A difference relative to Qwen includes the change of backbone and cannot, by itself, isolate an incremental memory benefit within the recipient. No additional recipient-side learning or model-specific memory-free result is assumed in this comparison.

E.4 Confusion Matrices, Subgroups, and Reliability

Figure 3 presents separate S1 and S2 matrices for each protocol. S1 compares normal and disease references; S2 compares non-tumor disease and tumor references within the reference-positive disease population. Each matrix includes a separate missing-output column. It is therefore a stage-specific classification view, not a joint three-class diagnosis matrix or a matrix restricted to successfully routed predictions.

For a stage-specific count matrix $M_{c,j}$ and reference support n_c, row normalization is

$$P_{c,j} = 100\,\frac{M_{c,j}}{n_c}, \qquad j \in \mathcal{C}_s \cup \{\perp\}, \quad n_c > 0. \tag{26}$$

The missing-output category $\perp$ follows the routing semantics in Appendix D: it includes abstention, a review decision, or stopping at an earlier stage. An S2 tumor reference predicted as normal remains

an S2 route miss rather than disappearing from that matrix. A normal reference remains outside the S2 population even if the model incorrectly predicts disease. This distinction preserves the separation between stage eligibility and predicted routing.

The diagonal expresses class recall, off-diagonal clinical columns express misclassification, and the missing-output column expresses unavailable stage decisions. Balanced accuracy averages the class recalls, whereas overall accuracy weights correct decisions by reference support. Coverage summarizes whether an eligible stage output was supplied. These views are complementary: high coverage can coexist with poor discrimination, and high recall for one class can coexist with errors or route misses in another. This is particularly relevant to the non-tumor/tumor asymmetry discussed in the main text. Neither the confusion matrices nor response coverage establishes probability calibration, clinical utility, or claim-level factuality.

The implementation retains per-class supports and prediction identifiers, allowing a subgroup metric to be defined by filtering paired predictions and references before applying the same scorer. Its denominator is then the eligible population within that subset, not a fraction of the full-cohort percentage. This describes the available scoring operation without adding unreported subgroup estimates. The reliability evidence considered here is the balanced accuracy, coverage, and stage-wise error structure reported in the main text.

E.5 Computational Cost and Memory Growth

The implementation distinguishes stored experience, active experience, and retrieved context. A snapshot contains stable and dynamic cards, including inactive dynamic cards retained for provenance. Active-memory capacity applies to retrievable dynamic cards, while the prompt budget applies only to the selected context. A growing audit archive therefore need not produce an equally growing inference prompt. Conversely, the same active-card count can correspond to different rules, lengths, or retrieval mixtures.

For active dynamic cards $\mathcal{E}_t^{\text{active}}$, the tracked state quantities include

$$N_t^{\text{active}} = |\mathcal{E}_t^{\text{active}}|, \qquad B_t^{\text{active}} = \sum_{e \in \mathcal{E}_t^{\text{active}}} \hat{\tau}(e), \tag{27}$$

where $\hat{\tau}(e)$ is the stored token estimate. Retrieval records separately retain selected stable/dynamic identifiers, estimated memory tokens, and whole-card omissions. Admission events retain action and reason codes, while capacity events identify demoted cards. Merging can accumulate supporting events without adding a new active card, and deactivation removes a card from retrieval without deleting its historical record. These distinctions make memory growth interpretable at the level of representation and access, rather than treating every recorded event as an additional prompt item.

Resource accounting uses physical provider requests rather than only successfully parsed predictions. Learning normally includes a prediction and a reflection request, whereas held-out inference includes prediction only; structural repair, transient retries, and loading requests can add to these totals. Successful provider responses supply input/output token usage. Failures are counted, and unavailable failed-request usage remains unknown instead of being treated as measured zero. This accounting separates the number of processed cases from the work performed by the inference service. It specifies how computational measurements can be attributed, without asserting an unreported latency reduction, cost advantage, or memory-quality gain.

F Qualitative Cases and Memory Audit Trails

Return to appendix contents

This section complements the aggregate results with an illustration of the hierarchical output contract and a description of the records supporting case-level inspection. The illustration explains how decisions are presented across stages. The audit representation connects an observed prediction to the memory actually supplied to the reasoner and to the learning events underlying that memory.

F.1 Simulated Question Answering and Result Display

Figure 4 presents three synthetic CBC examples corresponding to the principal decision paths: a normal pattern, a non-tumor disease pattern, and a possible tumor pattern. All patient profiles, measurements, and answers in the figure are illustrative; the answers were authored for presentation rather than measured from model executions. These examples are separate from the retrospective cohorts and contribute no observations to the quantitative evaluation.

Each panel pairs a question with CBC values and units, followed by a stage-specific answer and its evidential boundary. Panel (a) illustrates an S1 decision of no disease, after which S2 and S3 are not triggered. Panel (b) illustrates an S1 disease decision followed by an S2 non-tumor decision, which also terminates the route before tumor-group ranking. Panel (c) illustrates the disease–tumor-risk path and the resulting three-group S3 list. A downstream stage that is not triggered is distinct from a negative decision at that stage, consistent with the routing and missing-output definitions in Appendix D.

The highlighted measurements indicate which observations the authored answer emphasizes. For example, panel (b) connects its reactive-pattern description to WBC and absolute neutrophil values, whereas panel (c) emphasizes absolute lymphocytes, hemoglobin, and platelets. These annotations explain the intended relation between displayed evidence and the illustrative response; they do not validate a clinical diagnosis. In particular, the second and third entries in panel (c) demonstrate the three-position ranking format and are not established by the supplied CBC values. Neither the normal illustration nor the non-tumor illustration establishes the absence of malignancy.

The figure's prose is a reader-facing illustration of the task contract. The implemented reasoner instead returns schema-constrained decisions and at most three copied analyte/value observations, without a free-text rationale, as specified in Appendix C.3. Thus, the illustration clarifies stage activation and evidence presentation without implying that its explanatory prose, highlighted findings, or rankings were generated by the implemented model.

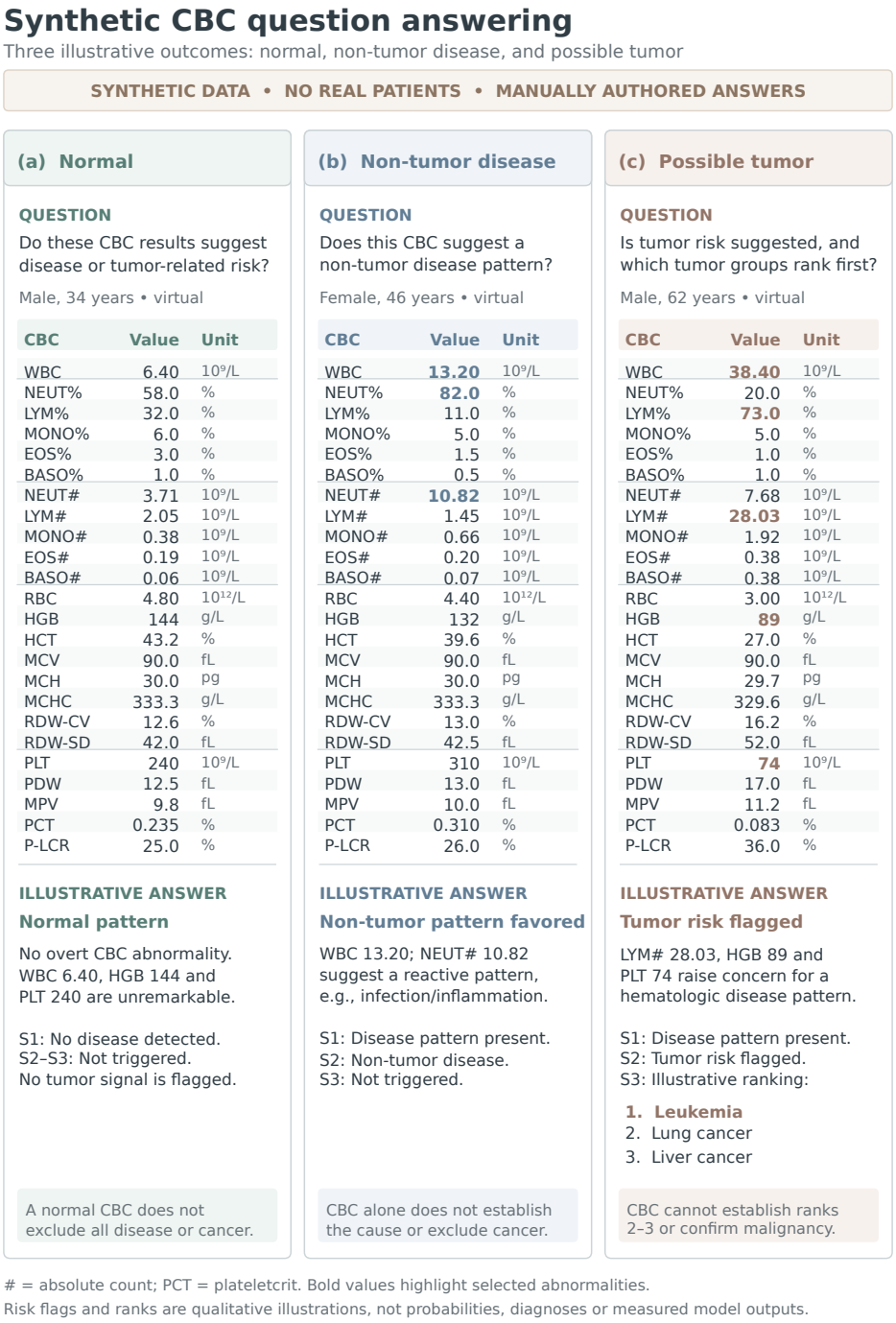


Figure 4: Synthetic illustration of hierarchical CBC question answering. Panels (a)–(c) show the normal, non-tumor disease, and possible tumor decision paths. All inputs are virtual and all answers are authored illustrations, not recorded model outputs. Highlighted values and the example ranking explain the presentation format; they are neither diagnostic evidence nor additional benchmark results.

F.2 Prediction Context and Memory Provenance

The implementation represents a prediction and its supporting memory history through linked records rather than a single generated narrative. Table 14 summarizes the relevant bindings. The case identifier connects a structured prediction to its retrieval record; the experiment manifest identifies the fixed memory resource. Card identifiers are resolved within that snapshot, so the inspected card text corresponds to the version available at prediction time rather than a later revision.

Table 14: Records supporting case-level inspection in the implemented memory workflow. Prediction records describe held-out inference; feedback, candidate, and admission records belong to earlier learning events.

Record	Binding and interpretive role
Case and prediction	Case identifier and structured stage outputs bind the decision to the model-visible input. References are paired separately for scoring.
Memory snapshot	Configuration and parent-state bindings, checkpoint, cards, catalog, and integrity hashes identify the memory resource used for inference.
Retrieval context	Selected sections and card identifiers, post-fitting retained identifiers, omitted cards, and token estimates identify the context actually supplied.
Learning episode	A restricted event record retains the learning input, reference, prediction, feedback, and retrieval metadata.
Candidate and admission	The event identifier links the reflection candidate to its admission action and reason codes, including recorded capacity demotions.
Card provenance	Source-event identifiers and card state connect reusable guidance to its contributing learning history and retrieval eligibility.

Context available at prediction time. Retrieval initially selects cards from the fixed memory, after which complete-prompt fitting may omit cards to satisfy the input limit. The prediction trace preserves both the pre-fitting selection and the retained identifiers, together with stable/dynamic counts and estimated memory tokens. A card present in the library, or selected before fitting, is therefore not sufficient evidence that its text reached the reasoner. Case-level interpretation depends on the retained context and the structured response for that same case.

History of a retrieved rule. For a dynamic card, the source-event identifiers connect its reusable content to restricted learning episodes. Each episode shares an identifier with its reflection candidate and the corresponding admission decision. The action and reason codes distinguish insertion, revision, consolidation, rejection, and conflict-related handling, while card activity records determine subsequent retrieval eligibility. Multiple supporting events can contribute to one retained card. These links describe how guidance entered the memory and which updates affected it; they do not establish that the guidance is medically correct or causally responsible for a particular prediction.

Separation of learning and evaluation. Held-out prediction loads a fixed parent snapshot, saves predictions and retrieval records, and checks that the bound memory remains unchanged. Its reference labels are not part of the reasoner's input, and it has no reflection or admission stage. Consequently, a retrieved card's learning history belongs to earlier learning cases, not to feedback from the held-out case being explained. The synthetic examples in Fig. 4 contain no such execution history and are used only to illustrate the output contract. No case-specific improvement, clinical adjudication, or memory-attribution result is inferred from them.

G Limitations and Future Work

Return to appendix contents

The framework enables a frozen language model to reuse experience through explicit, revisable memory. Separating candidate generation from controlled admission, protecting stable knowledge, and retaining provenance provide a structured basis for examining how feedback shapes subsequent context. The present results support this approach to retrospective risk-oriented reasoning. The following considerations clarify the scope of that evidence and identify directions for extending the method.

G.1 LIMITATIONS

- **Scope of the empirical evidence.** The reported improvements are most consistent for disease detection and tumor-group prioritization, while tumor-risk assessment remains sensitive to the metric and protocol. For example, RAG has higher S2 accuracy on MSA, and ACE has higher S2 F1 on HMT. The MSA S3 margins over RAG are small, which favors interpreting the ranking results as competitive performance rather than a substantial separation on that protocol. The checkpoint curves complement these comparisons by showing performance as memory evolves; additional uncertainty and order-sensitivity analyses would further characterize the stability of the observed patterns.
- **Evidence scope and task formulation.** CBC measurements and limited accompanying information provide a focused setting for studying memory-assisted reasoning. The three-stage formulation and eight-group tumor vocabulary make evaluation explicit, while defining the scope of the resulting predictions. In particular, S3 measures prioritization within the specified label space rather than a calibrated differential diagnosis over all diseases. The outputs are therefore best interpreted as risk-oriented assessments under the available evidence. Broader clinical information would help resolve distinctions that remain ambiguous from CBC findings alone.
- **Population and source dependence.** MSA and HMT examine memory adaptation under complementary source compositions. Normal controls originate from the hospital cohort, and the MIMIC component consists of selected inpatient admissions. HMT evaluates hospital-initialized memory after MIMIC updates on a mixed-source test cohort. This design provides evidence for adaptation across the studied source transition; its applicability to other institutions, patient groups, and measurement conventions remains an important area for further evaluation.
- **Memory validity and mechanistic interpretation.** Deterministic admission rules and linked provenance make memory changes inspectable and repeatable. The usefulness of an admitted rule nevertheless depends on the quality of feedback and the scope of its generalization. Provenance supports examination of these dependencies, while medical validation of reusable guidance remains complementary to structural checks. The component comparisons also reflect the implemented configurations, including their consolidation and retrieval allocations. More targeted controls would help distinguish the contributions of individual operations to the overall behavior.

G.2 FUTURE WORK

- **Broader clinical evidence and validation.** A natural extension is to incorporate additional clinical context, such as longitudinal measurements, clinical notes, or imaging findings, while preserving the separation between reusable guidance and case-specific evidence. Source-stratified analyses, independent institutional cohorts, and prospective clinician-supervised studies could clarify how the approach transfers to broader clinical settings and supports interpretation under richer evidence.
- **Feedback quality and memory refinement.** The recorded candidates, admission decisions, and card histories provide a basis for studying sensitivity to feedback quality, case order, and conflicting experience. Expert review of selected rules and emitted observations could complement the current task metrics and guide refinement of applicability conditions. Further work could also examine how confidence and review routing relate to uncertainty, with the aim of making memory reuse more responsive to the evidence available for each case.
- **Cross-backbone reuse and resource allocation.** The recipient-model results illustrate the portability of the external-memory representation. Paired comparisons with and without transferred memory for each recipient would help quantify its incremental contribution across backbones. The configurable context interface and automatic prompt fitting also support further study of task performance alongside realized context length, request usage, and latency. Such analyses could guide the choice of memory budgets for different models and computational settings beyond the two profiles used in the present comparison.

- **Responsible research translation and broader impact.** Explicit memory and traceable revisions may facilitate review, correction, and collaboration in research on medical reasoning agents. Future translation can build on these properties through clinician involvement, appropriate access controls for episode records and derived memory, and evaluation across patient groups. Clear communication of the risk-oriented output scope would help users interpret predictions alongside other clinical evidence. These directions aim to preserve the method's inspectability while assessing its practical value in supervised decision-support settings.