# WHEN SHOULD RAG WRITE? SELECTIVE MEMORY FOR CONTINUAL MEDICAL QA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Medical question answering systems can learn from answer feedback, but a source-supported claim may still be inapplicable to a later patient. We introduce EvoMedAgent, a selective admission policy for graph memory with a frozen generator. The policy couples three mechanisms: executable clinical scope that must pass before retrieval ranking, source-family calibration of assertion support, and joint admission checks for source conditions and unresolved conflicts. The policy requires patient-bound observations to remain patient-specific and general assertions to retain reuse conditions. Within a test-then-learn protocol, accepted assertions become visible to later questions. Evaluation spans six frozen medical QA benchmarks, paired online streams, and transfer to disjoint EHRNoteQA patients after memory is frozen. Relative to no update, the full policy improves MedQA AULC by 5.3 points and EHRNoteQA transfer accuracy by 3.5 points. Matched-editor comparisons and component ablations establish the contribution of admission checks, while feedback-schedule controls examine evidence availability. Together, these results show that selective admission improves later predictions with a frozen generator and explicit checks on the reuse conditions of accepted evidence.

## 1 INTRODUCTION

A medical QA system can retain answer feedback without retraining its generator, but even a source-supported lesson may not apply to a later patient. Patient characteristics and clinical context constrain reuse: if these applicability conditions are lost during writing, semantic retrieval can surface an overgeneralized claim despite a close textual match. Medical memory must therefore preserve both the evidence for a claim and the conditions under which it applies to a later patient.

Writable-memory systems already separate prediction from subsequent context updates. Dynamic Cheatsheet and ACE use this ordering, while ReasoningBank distills reusable lessons from completed trajectories (Suzgun et al., 2026; Zhang et al., 2026; Ouyang et al., 2026). This ordering allows feedback to improve later predictions while keeping it out of the current answer. Medical reuse additionally requires preserving a lesson's source support and clinical conditions throughout the update. We therefore adopt this test-then-learn protocol with fixed generator weights and answer-supervised updates, asking which medical claims should enter memory and under what conditions they can be reused.

EvoMedAgent makes source support and clinical applicability joint requirements for memory admission and reuse (Figure 1). **Medical eligibility:** assertions retain patient identity and physiological, clinical, and medication/intervention scope; incompatible or unresolved constrained attributes exclude candidates before ranking. **Source-calibrated support:** calibration on disjoint source families maps verifier scores to agreement with a separate reference model's source-based judgments, determining group and candidate eligibility. **Joint admission:** support, source/schema validity, scope preservation, and conflict checks govern publication; unresolved contradictions remain unretrievable. Conditions preserved at writing are enforced at reading.

We evaluate whether this policy improves subsequent answers, which admission checks contribute, and whether learned memory transfers to unseen patients. Twenty-five online trajectories pair methods by item order and run index, with shared initialization, generator, and budgets. Area under the learning curve (AULC) summarizes adaptation, while copied-state controls and controlled masking (CM)



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

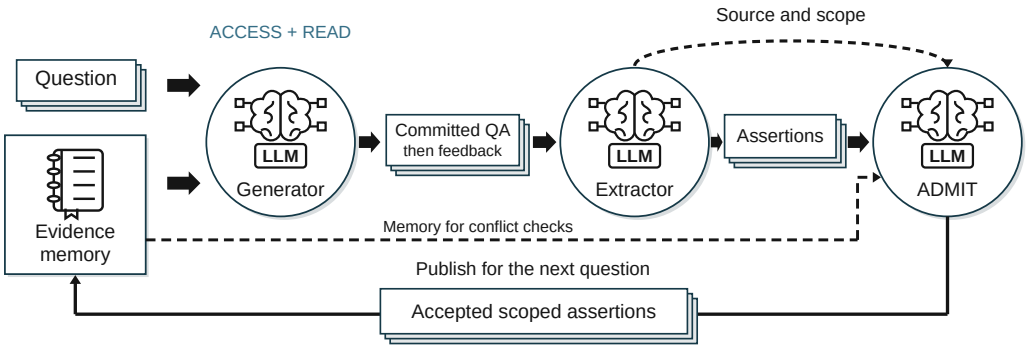


Figure 1: **Selective admission for medical memory.** Access and Read supply the frozen generator with eligible evidence. After commitment and feedback, the extractor proposes scoped assertions; Admit combines model-based support with calibrated eligibility checks before publishing them for the next question.

examine admission and evidence availability. The policy improves online performance over matched editors and transfers to disjoint EHRNoteQA patients with memory frozen, linking selective writing to gains in subsequent answers.

## 2 Related Work

**Medical models and evaluation.** MIRAGE compares medical corpora, retrievers, and generators under shared questions (Xiong et al., 2024); EHRNoteQA uses discharge summaries (Kweon et al., 2024). ReasonMed develops medical reasoning data, while Med-PaLM 2 and MedHELM examine medical capability and broader evaluation criteria (Sun et al., 2025; Singhal et al., 2025; Bedi et al., 2026). We combine frozen benchmarks and sequential memory evaluation with task-specific formats and shared references, distinguishing retrieval quality from adaptation.

**Graph retrieval and non-parametric memory.** HippoRAG 2, MedGraphRAG, BYOKG-RAG, and AMG-RAG use entity linking, propagation, and domain structure (Gutiérrez et al., 2025; Wu et al., 2025; Mavromatis et al., 2025; Rezaei et al., 2025). CatRAG and A2RAG adapt traversal or escalation (Lau et al., 2026; Liu et al., 2026b). HippoRAG 2 studies continual memory; AMG-RAG combines continuing graph construction and confidence-guided traversal. Building on structured retrieval, we address the complementary problem of governing post-answer assertion admission and reuse through source calibration and medical scope.

**Dynamic editing, uncertainty, and selective compute.** KEDKG handles conflicting facts through graph editing (Lu et al., 2025). Adaptive-RAG selects retrieval strategies by question complexity; confidence signals and selective reasoning allocate additional computation (Jeong et al., 2024; Chuang et al., 2025; Wang et al., 2026; Zhan et al., 2026). C-RAG bounds generation risk, while SURE-RAG and TCR examine evidence sufficiency or conflict at answer time (Kang et al., 2024; Qiu et al., 2026; Ye et al., 2026). Our policy extends these checks to evidence retained after feedback, when an assertion may affect many later answers. Paired update trajectories and accuracy–cost comparisons distinguish memory adaptation from selective computation.

**Continual learning and writable memory.** Parameter-centric methods manage interference through weight updates (Li et al., 2024; Cheng et al., 2025; Shi et al., 2025); CMT uses compressed latent memory (Li et al., 2025), and A-Mem revises linked notes (Xu et al., 2025). Dynamic Cheatsheet's DC-Cu and ACE answer before updating memory or context; ACE also studies online adaptation with ground-truth labels (Suzgun et al., 2026; Zhang et al., 2026). ReasoningBank extracts reusable strategies from completed trajectories judged successful or unsuccessful by a model (Ouyang et al., 2026). Building on these update protocols, EvoMedAgent adds source-family calibration, executable medical scope, and conflict isolation. Appendix M.1 compares AMG-RAG and ACE policies; matched experiments cover graph editors and DC-Cu.

# 3 METHOD

## 3.1 OVERVIEW: ADMISSION AND CONDITIONAL REUSE

EvoMedAgent keeps the generator fixed and combines immutable background knowledge with accepted dynamic assertions carrying sources and scope (Figure 2). To separate prediction from learning, each step freezes visible memory: ACCESS routes, READ selects evidence, and the generator commits its answer and evidence IDs before feedback. ADMIT then checks lessons from the revealed answer; accepted assertions become visible at the next question, with each visibility change logged. Frozen evaluation stops after scoring.

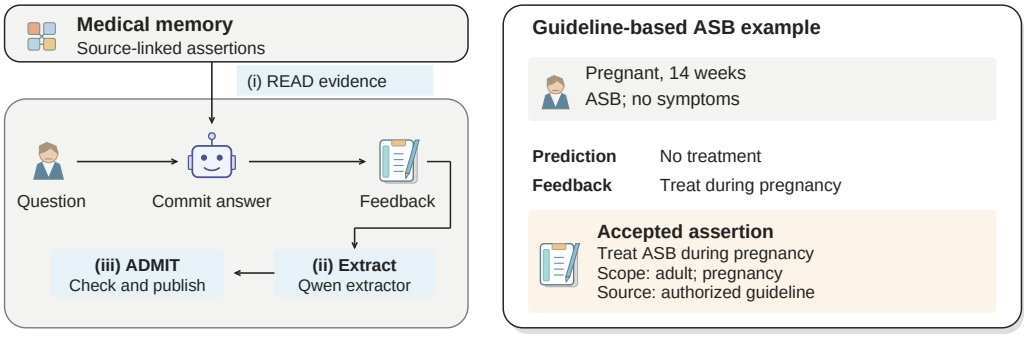


Figure 2: **Clinical scope governs assertion reuse.** Memory retrieval precedes answering; extraction and admission follow feedback. In this asymptomatic bacteriuria (ASB) example, feedback includes a guideline recommendation; the assertion retains its source and pregnancy scope (Nicolle et al., 2019).

The policy requires clinical qualifiers and certainty at writing (Figure 3). Stored patient-bound assertions are retrievable only for that patient; general assertions use null identifiers and scope matching, including MedQA/MedMCQA and guideline lessons. Because these checks operate on stored attributes, extraction errors can still leave claims overgeneralized or missing necessary conditions.

## 3.2 ACCESS: DOES THIS QUESTION NEED RETRIEVAL?

ACCESS combines deterministic uncertainty $H(q)$ and linked-concept density $D(q)$ probes. For multiple-choice questions, $H(q)$ is option-probability entropy divided by $\log|\text{options}|$; for open-ended questions, it averages token entropy over a greedy probe of at most 32 steps. Density counts concept mentions with weights 1 for linked, 0.5 for unresolved, and 0 for nonconceptual mentions. Construction-data percentile maps within question-length buckets yield $\widetilde{H}(q)$ and $\widetilde{D}(q)$, with zero density mapped to zero. Appendix G specifies normalization, bucket merging, and probe execution. Invalid uncertainty probes route directly to Deep. The routing score is

$$C(q) = w_H \widetilde{H}(q) + (1 - w_H)\widetilde{D}(q). \tag{1}$$

Choose Deep for $C(q) > \tau$ and Fast otherwise. Non-evaluation data fix temperature, maps, weight, and threshold before evaluation; frozen repetitions share one route. If no configuration qualifies for a backbone–dataset pair, log an Always-Deep fallback with the same generator, retrieval settings, and decoding parameters. Hosted models without full logits use a separately selected density-only rule. Appendix G gives the selection criterion and operating points.

## 3.3 READ: WHICH EVIDENCE FITS THIS QUESTION?

Fast supplies no retrieved evidence. Deep searches background and currently visible dynamic assertions. It retrieves at most 100 approximate nearest-neighbor candidates and checks visibility and the dynamic-edge similarity threshold. READ keeps patient-bound candidates only when their

identifier exactly matches the resolved current patient; null-ID candidates skip this check. READ then checks physiological state (age, sex, pregnancy/lactation, weight, renal function), clinical context (conditions, allergies, care setting), and medication/immunization status. All declared constraints must pass; unresolved constrained fields are incompatible (Appendix H.1). Identity and scope checks precede ranking and cannot be overridden by textual similarity.

Remaining candidates receive the weighted score

$$s(q, e) = 0.80\ \mathrm{sim}(q, e) + 0.20\, r(e) - 0.05\, d(q, e). \tag{2}$$

Similarity is the MedCPT cosine score clipped to $[0, 1]$. Reliability $r(e)$ uses calibrated verifier scores for dynamic assertions and frozen source-stratum support estimates for static edges. The mismatch $d(q, e)$ is zero if either endpoint matches a linked query entity and one otherwise. Evidence retains its certainty in the answer prompt.

Candidates are scanned in descending score order, with ties broken by reliability then canonical edge key. The pack contains at most 20 edges, including at most five dynamic edges, within 1,024 evidence tokens. Shared caps keep answer budgets comparable as memory grows (Appendix H.1). Ranking cannot recover clinical conditions omitted from a stored claim, so scope preservation at admission remains necessary.

The implementation uses UMLS/SapBERT for entity linking, PrimeKG for background edges, and MedCPT for query–triple similarity (Liu et al., 2021; Chandak et al., 2023; Jin et al., 2023). Appendices B and D specify assertion fields and the generator's evidence format.

### 3.4 ADMIT: WHICH LESSONS SHOULD BE STORED?

After scoring, a frozen extractor receives the completed question stem and released feedback answer text to propose scoped assertions. Its input omits the option list and the model's own answer, so unselected alternatives cannot serve as extraction evidence. For each candidate $e$,

$$\mathrm{keep}(e) = \mathbf{1}[r(e) \geq 0.70\ \wedge\ \mathrm{valid}(e)\ \wedge\ \mathrm{scoped}(e)\ \wedge\ \neg\,\mathrm{conflict}(e)]\,. \tag{3}$$

Here, $r(e)$ is calibrated source support; validity requires an approved calibration group, authorized source, and valid schema, entity links, and times. A separate MedGemma call compares the full stem and released feedback with the candidate and its scope; $\mathrm{scoped}(e)$ requires a judgment of no omitted or broadened source conditions (Appendix D). The conflict flag marks unresolved contradictions. Accepted assertions become visible at the next question; duplicates merge sources. Parser exceptions can bypass source-span validation and allow assertions with unverifiable provenance (Appendix E.1).

To make verifier scores usable for admission, we calibrate their agreement with source-based judgments. MedGemma-27B-Text-IT supplies support scores. Before evaluation, frozen Selene Mini from another model family labels source-family-disjoint calibration candidates using sources alone, without verifier scores, rationales, method names, or admission decisions. Calibration maps scores to reference agreement. Appendix D gives prompts; Figure 3 details the admission checks and the resulting patient-bound record.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

We first evaluate answer quality with memory held fixed, then measure learning from answer feedback and transfer to unseen patients. Matched admission, retrieval, and routing comparisons examine the mechanisms behind these outcomes. Online methods maintain independent memories, including each trajectory's effects of earlier updates on later predictions.

**Datasets and protocols.** Construction and selection use half of each training split exceeding 3,000 items and all items otherwise, with no evaluation overlap. These data fix temperature, feature maps, routing, and retrieval. Frozen evaluation uses five answers per item; paired analyses retain complete five-generation blocks across compared methods.

MedQA and both MMLU tasks use official splits; MedMCQA evaluates on official validation and builds memory from filtered training IDs (Appendix A.2). EHRNoteQA v1.0.0 uses a patient-disjoint

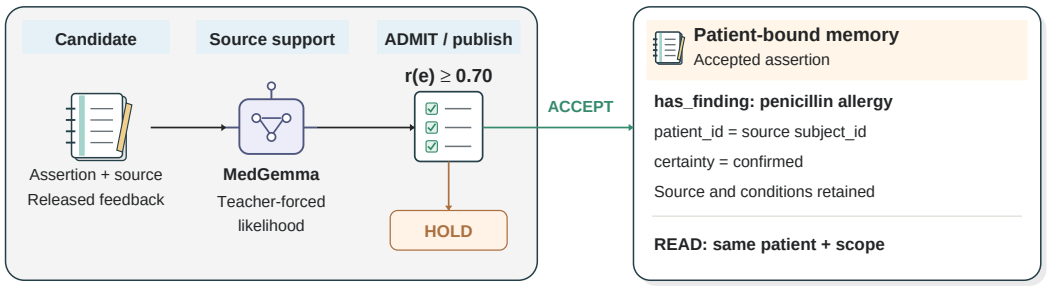


Figure 3: **Admission checks the conditions of a claim.** Source support and calibrated eligibility checks govern publication. The allergy assertion retains its patient identifier and clinical scope; Read checks both before ranking.

462/500 construction/evaluation split of 962 records. PubMedQA uses separate training, validation, and test partitions of PQA-L, with tuning restricted to validation data (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019; Hendrycks et al., 2021; Kweon et al., 2024). Shared initial memory excludes evaluation membership information and is filtered for leakage (Appendices F and E.2); EHRNoteQA transfer starts with static-only memory (Appendix N).

Online MedQA and MedMCQA each use 25 paired trajectories: independently maintained memories, uniformly shuffled fixed pools, and one sampled answer per item. Shared order seeds 41001–41025 and generation-run seeds 51001–51025 pair trajectories; request seeds also include method identity. Post-scoring feedback supports immediate admission, a 100-step delay, past-feedback permutation, and controls matching final update counts. Permutation moves complete question–answer pairs; positions index presentation order.

**Matched baselines.** Baselines include direct and chain-of-thought prompting, Vector RAG, Adaptive-RAG, Selective CoT, MedGraphRAG, HippoRAG 2, and Always-Deep (Jeong et al., 2024; Zhan et al., 2026; Wu et al., 2025; Gutiérrez et al., 2025); online adaptations include incremental HippoRAG 2, KEDKG, and confidence-aware KG (Lu et al., 2025; Takahashi et al., 2025). Methods share the generator, source pool, caps, decoding, and hardware; online editors also share answer feedback. Passage baselines use source-coherent text from graph methods' records. Appendix E details adaptations and paper-specified controls.

**Backbones and implementation.** Nine matched systems use Qwen3.5-9B; sensitivity covers six open checkpoints (0.8B–27B) and three hosted GPT-5.6 variants (Sellergren et al., 2025; Liu et al., 2026a; Qwen Team, 2026a;b; OpenAI, 2026). Routing uses full-logit H+D locally and density only for APIs. Local models run on one A800 80GB with shared stochastic answer decoding; probes, extraction, and judging are deterministic. Appendices D and K specify sampling parameters, seeds, output caps, model revisions, templates, software, and API interleaving.

**Evaluation metrics.** Frozen accuracy averages evaluable answers within each item, then weights items equally; each item requests five answers. The deterministic, method-masked Selene Mini judge counts EHR rubric scores of four or five out of five as correct (Alexandru et al., 2025). AULC averages trailing-window accuracy (up to 100 predictions); Final uses the last 20%. Routing reports latency and escalation; evidence precision/recall and Read/Write ECE use offline Selene references (scoring: Appendices A and D).

**Statistical analysis.** Construction/selection also fix checkpoints, sampling, evaluators, the EHR rubric, calibration, and inference. Frozen comparisons pair methods within each item/patient; online comparisons pair methods within each trajectory. Online intervals reflect order and answer sampling for the fixed pool/configuration (Holm families: Appendix E.3).

Table 1: **Frozen medical QA on Qwen3.5-9B.** Each item requests five answers; accuracy (%) weights item means equally after exclusions. MedQA uses 1,273 items; online MedQA uses 1,000 items and 25 trajectories (Table 2).

| | MedQA | MedMCQA | PubMedQA | MMLU-Gene | MMLU-Clinic | EHRNoteQA |
|---|---|---|---|---|---|---|
| *Prompting and selective reasoning* | | | | | | |
| Direct | 70 | 64 | 75.4 | 79 | 78.6 | 65.8 |
| Always-CoT | 70.3 | 64.5 | 75.7 | 79.2 | 79 | 66.1 |
| Selective CoT | 70.9 | 65.1 | 76.3 | 80 | 79.2 | 66.7 |
| *Text retrieval and adaptive RAG* | | | | | | |
| Vector RAG | 72.7 | 67 | 77.6 | 82.8 | 81.9 | 68.7 |
| Adaptive-RAG | 73.8 | 68.2 | 78.5 | 83.2 | 82.9 | 69.7 |
| *Graph-grounded retrieval* | | | | | | |
| MedGraphRAG (matched) | 79.2 | 73.8 | 84 | 87.8 | 88.5 | 75.3 |
| HippoRAG 2 | 80.1 | 74.6 | 84.9 | 89 | 89.6 | 76.3 |
| Always-Deep | 80.5 | 75 | 85.4 | 89.4 | 90 | 76.6 |
| **EvoMedAgent** | **82.1** | **77** | **86.6** | **92** | **91.7** | **78.4** |

## 4.2 Main Results

**Frozen medical QA.** Under the shared Qwen3.5-9B backbone, EvoMedAgent achieves the highest accuracy estimate on all six datasets in Table 1. Gains range from 1.2–2.6 points over Always-Deep and 1.7–3.0 over HippoRAG 2. This comparison measures answer quality with memory held fixed. The online evaluation below then measures learning from subsequent feedback.

Table 2: **Online learning.** Twenty-five paired trajectories; AULC/Final use trailing-100/last-20% windows. Panel A: EvoMedAgent minus editor, marginal 95% CIs. CM restores initial memory over 100 steps and varies feedback release.

**A. MedQA: matched editors**

| Editor | Final | AULC | Δ AULC 95% CI |
|---|---|---|---|
| Inc. HippoRAG 2 | 83.3 | 82.7 | +4.1 [+3.3, +4.8] |
| KEDKG | 84.7 | 84.2 | +2.6 [+1.8, +3.1] |
| Confidence KG | 83.8 | 83.7 | +3.1 [+2.5, +4] |
| Supervised DC-Cu | 86.32 | 85.56 | +1.24 [+0.99, +1.49] |
| **EvoMedAgent** | 87.4 | 86.8 | – |

**B. Update schedules**

| | MedQA | | MedMCQA | |
|---|---|---|---|---|
| Policy | Final ↑ | AULC ↑ | Final ↑ | AULC ↑ |
| **No update** | 82.00 | 81.50 | 76.5 | 76.2 |
| **Immediate** | **87.4** | **86.8** | **80.2** | **79.8** |
| CM immediate | 87.1 | 86.5 | 79.6 | 79.4 |
| CM delayed | 86.1 | 85.5 | 77.9 | 77.6 |
| CM permuted | 86.3 | 85.7 | 78.4 | 77.9 |

**C. MedQA endpoints**

| Policy | Acc. (%) | AULC | Final (%) |
|---|---|---|---|
| No update | 81.8 | 81.50 | 82.00 |
| Immediate | 87.1 | 86.8 | 87.4 |

**Online learning.** With the same generator and retrieval budgets, selective retention of answer feedback improves online performance (Table 2). Calibrated admission gains 5.3 AULC and 5.4 Final points over no update (Panel C). Among matched editors (Panel A), its paired AULC gain over KEDKG is 2.6 points (95% CI [+1.8, +3.1]). It also exceeds supervised DC-Cu by 1.24 AULC points (95% CI $[+0.99, +1.49]$) and 1.08 Final points (95% CI $[+0.88, +1.3]$), with positive Final differences in 25/25 trajectories. Each editor maintains its own evolving memory from the same feedback stream. The CM schedules (Panel B) vary evidence availability: CM immediate exceeds CM delayed by 1.00 AULC points on MedQA and 1.8 on MedMCQA, and exceeds CM permuted by 0.80 and 1.5, respectively. CM restores initial memory over 100 steps and varies feedback release; no update retains initial memory in full. Their intermediate evidence therefore differs. Panel C reports whole-stream accuracy alongside the windowed endpoints, and Table 3 separates the effects of admission checks over complete histories.

**Learning beyond the initial window.** We separate burn-in from later predictions to examine performance after memory restoration. CM masking gives first-100 MedQA AULC of 70.0–70.7; after restoration, AULC is 88.26 for CM immediate, 87.22 for CM delayed, and 87.44 for CM permuted. Each trajectory ends with 6082 visible transactions and 2147 merged edges. The corresponding MedMCQA scores are 79.8, 78.0, and 78.3 (Appendix M). This schedule comparison combines evidence availability, content, and update timing.

### 4.3 Ablation and Efficiency Analysis

Table 3: **Online MedQA write comparisons.** 25 paired 1,000-item trajectories per comparison. Panels use their own paired Full references and shared ACCESS/READ settings. A: four-test AULC Holm family. B: marginal paired 95% CIs. Scope veto denotes the ADMIT call; source-auth gate covers authorization, metadata, and source–relation eligibility (Appendix E.1).

**A. Gate ablations:** $\Delta = \text{variant} - \text{Full}$

| Variant | AULC | Final | Final− AULC | Δ AULC | Δ Final | AULC 95% CI | Raw $p$ | Holm $p$ |
|---|---|---|---|---|---|---|---|---|
| Full | 86.8 | 87.4 | 0.6 | | | Reference | | |
| Raw confidence | 86.09 | 87.04 | 0.95 | −0.71 | −0.36 | $[-1.03, -0.39]$ | 0.0130 | 0.013 |
| No conflict isolation | 85.3 | 85.46 | 0.16 | −1.5 | −1.94 | $[-1.88, -1.12]$ | 0.0010 | 0.002 |
| No ADMIT scope veto | 84.73 | 86.26 | 1.53 | −2.07 | −1.14 | $[-2.45, -1.69]$ | $< 0.00033$ | $< 0.001$ |
| No source-auth gate | 83.38 | 83.8 | 0.42 | −3.42 | −3.6 | $[-3.85, -2.99]$ | $< 0.00025$ | $< 0.001$ |

**B. Budget-matched comparison:** $\Delta = \text{Full} - \text{variant}$

| Variant | AULC | Final | Δ AULC [95% CI] | Δ Final [95% CI] |
|---|---|---|---|---|
| Full | 86.7681 | 87.08 | Reference | |
| Generic recheck | 85.7044 | 86.46 | $+1.0637$ $[0.6060, 1.5480]$ | $+0.62$ $[-0.02, 1.26]$ |
| No ADMIT scope veto | 84.9006 | 85.86 | $+1.8675$ $[1.4132, 2.3264]$ | $+1.22$ $[0.72, 1.68]$ |

**Memory admission.** Table 3 compares write policies under shared ACCESS/READ settings; raw confidence changes admission scores only. With matched recheck call and token caps, Full exceeds Generic by 1.0637 AULC points (95% CI $[0.6060, 1.5480]$) and the scope-veto ablation by 1.8675 ($[1.4132, 2.3264]$). The Full–Generic Final difference is $+0.62$ points, with an interval including zero ($[-0.02, 1.26]$). For the separate scope-check decision, 100 paired tests give 4 versus 20 erroneous acceptances ($p = 0.000145$), with intact retention of 98 versus 95. In the clustering-policy comparison (Table 10), Full–Generic AULC gaps are 0.88 under source-unit and 0.81 under document-cluster policy ($[0.42, 1.20]$ for the latter). Frozen shared-proposal controls (Table 5) and assertion perturbations (Figure 7) complement these online comparisons.

**Routing efficiency.** Selective access routes 46.7% of MedQA queries to Deep (Table 4A). Relative to Always-Deep, median answer-path latency falls by 27.7%, from 1.77 to 1.28 seconds, while point accuracy increases by 1.6 points. H+D also exceeds density-only routing by 7.1 accuracy points (95% CI [+5.8, +8.4]; $p_{\text{Holm}} < 0.0001$). These measurements cover routing, retrieval, generation, and commitment; post-answer extraction, verification, and indexing are accounted for separately.

**Retrieval quality.** With identical evidence caps, full retrieval reaches 82.1% accuracy and 0.781 pooled-reference Recall@20 (Table 4B). HippoRAG 2 reaches 80.1% and 0.741, while confidence-aware KG reaches 80.3% and 0.752. The paired accuracy and recall comparisons each pass their six-test Holm correction. Table 5 also shows that combining both memories reaches 82.1% accuracy, compared with 79.9% for dynamic-only and 76.6% for static-only memory. Together, these results indicate that admitted assertions and background knowledge provide complementary evidence within the same prompt budget.

Table 4: **Access and retrieval on MedQA.** Shared caps; 27.7% lower answer-path median latency. A/B: separate timing runs (0.01 s batch variation). Esc.: native path escalation.

**A. Selective access**

| Method | Acc. ↑ | Esc. % | p50 / p95 s |
|---|---|---|---|
| Always-Fast | 70 | 0 | 0.92 / 1.31 |
| Always-Deep | 80.5 | 100 | 1.77 / 2.86 |
| Selective CoT | 70.9 | 43.1 | 1.31 / 2.04 |
| Adaptive-RAG | 73.8 | 49.5 | 1.39 / 2.18 |
| Entropy only | 74.7 | 51 | 1.34 / 2.16 |
| Density only | 75 | 48.3 | 1.29 / 2.07 |
| **EvoMedAgent** | 82.1 | 46.7 | 1.28 / 2.03 |

**B. Matched retrieval**

| Retriever | Acc. ↑ | Pooled R@20 | p50 s ↓ |
|---|---|---|---|
| Dense retrieval | 77.6 | 0.612 | 0.74 |
| Query–triple only | 78.4 | 0.681 | 0.88 |
| PPR only | 78.9 | 0.704 | 0.96 |
| Query–triple + PPR | 79.6 | 0.728 | 1.08 |
| HippoRAG 2 | 80.1 | 0.741 | 1.16 |
| Confidence KG | 80.3 | 0.752 | 1.21 |
| **EvoMedAgent** | 82.1 | 0.781 | 1.27 |

Table 5: **Historical-memory components.** Frozen MedQA, 1,273 items, five answers/item. $\Delta$: Full minus variant; marginal paired 95% CIs. All eight contrasts pass their Holm family.

| Variant | Acc. | $\Delta$ Acc. [95% CI] |
|---|---|---|
| Static only | 76.6 | +5.5 $[+4, +7]$ |
| Dynamic only | 79.9 | +2.2 $[+0.9, +3.5]$ |
| No semantic rerank | 79 | +3.1 $[+1.8, +4.4]$ |
| No confidence rerank | 79.5 | +2.6 $[+1.3, +3.9]$ |
| Dense; no links | 77.6 | +4.5 $[+3.2, +5.8]$ |
| Raw confidence | 80.1 | +2 $[+0.7, +3.3]$ |
| No source-auth gate | 79.2 | +2.9 $[+1.5, +4.3]$ |
| No conflict isolation | 79.6 | +2.5 $[+1.1, +3.9]$ |
| **Full** | **82.1** | – |

On the same frozen evaluation items, write variants rebuild memory from shared historical proposals; READ variants change evidence selection. Current feedback is excluded. All variants use $K = 20$, $K_d = 5$, and 1,024 evidence tokens.

### 4.4 Forward Transfer to Disjoint Patients

To test cross-patient reuse, memory learned on 462 EHRNoteQA patients is frozen for 500 disjoint patients. Each of 25 trajectories independently permutes both streams, starts from static-only memory, and disables downstream writes. Transfer and masking use Always-Deep; five answers per patient share pre-feedback memory (Figure 4; protocol: Appendix N). Full gains 3.5 accuracy points over no update on method-specific evaluable sets (95% CI $[2.8, 4.2]$). Masking tests reliance on query-relevant assertions in 2,376 common-evaluable trajectory–patient requests: 19.008% of all requests, covering 475/500 patients. Within this subset, matched random masking exceeds target masking by 2.2566 points under J0 ($[1.9063, 2.6207]$) and 2.6563 under J1 ($[2.1124, 3.1903]$); J1's Full–Random interval includes zero.

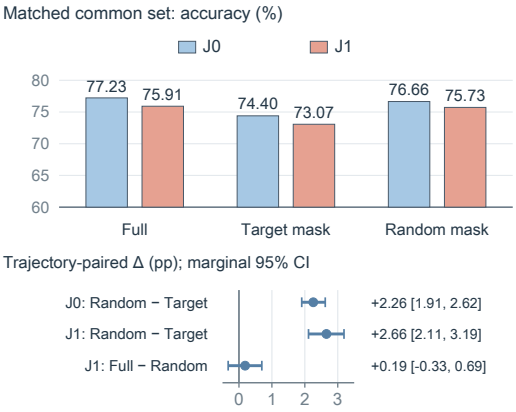


Figure 4: **Patient-disjoint EHR transfer.** Memory from 462 patients, frozen for 500 others. Top: J0/J1 accuracy on the common matched set. Bottom: selected masking effects with marginal paired 95% CIs (Appendix N).

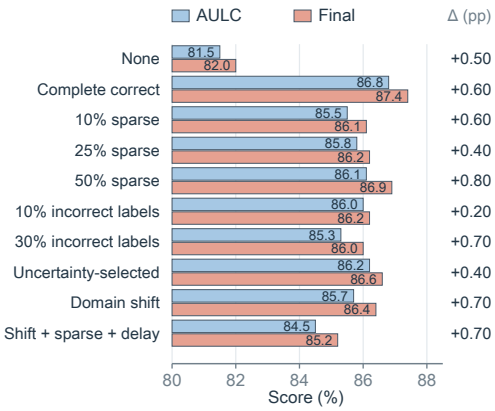


Figure 5: **Feedback availability and noise.** MedQA: 1,000 items, 25 trajectories, last-200 Final; all conditions score the full pool. $\Delta$ = Final − AULC. Combined: 50% feedback, 240-position delay; rules in Appendix M.1.

In a 7,200-entry context subset, source/answer-ID linkage is 98.38%/98.97%. Model classification with partial human correction assigns 100 sampled entries to 92 supported, five erroneous, and three unresolved (Appendix N.1).

### 4.5 Feedback Availability and Label Noise

Figure 5 compares sparse, delayed, and incorrect feedback on a common scored pool. Bernoulli masks are shared across methods; incorrect labels replace feedback answer text while scoring retains the

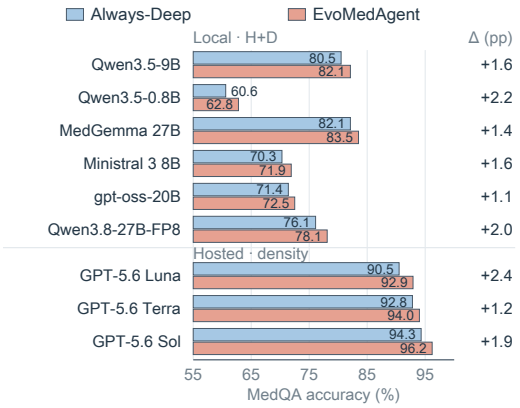


Figure 6: **Matched-backbone MedQA.** Frozen accuracy (%); Δ is relative to Always-Deep. Nine matched backbones cover local H+D and hosted density-only routing.

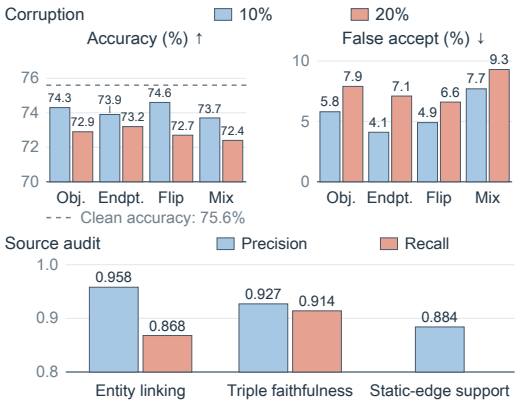


Figure 7: **Corruption and source audit.** Obj./Endpt.: object replacement/endpoint redirection; Flip/Mix: polarity flip/equal mixture. FA (%): retrieval-eligible share of model-rated unsupported injections. Lower: source-reference precision/recall.

original truth. Increasing feedback from 10% to 50% raises AULC from 85.5 to 86.1; 30% incorrect labels gives 86.0% Final. A constructed shift places two 250-item domains in each half, yielding AULC/Final of 85.7/86.4. With that same order, 50% feedback and a 240-position delay reduce the endpoints to 84.5/85.2. Appendix M.1 specifies masks, source reconstruction, and changed-domain scoring for this combined intervention.

### 4.6 Backbone and Assertion-noise Sensitivity

Figure 6 shows matched MedQA gains of 1.1–2.4 points over Always-Deep, using full logits for local H+D and density-only access for hosted models. Paired online comparisons use Qwen3.5-9B.

Figure 7 tests admission under assertion corruption with re-run verification, fixed pairings/source spans, and five greedy replicates on 500 Deep items. All four corruptions reduce accuracy and increase false acceptance with noise; clean 75.6% refers only to this subset. Appendix L specifies the interventions and source-audit denominators.

## 5 Discussion and Broader Impact

Online and shared-proposal comparisons quantify write-side effects (Tables 2, 3, and 5); patient-disjoint evaluation tests frozen-memory reuse within one dataset.

Feedback availability also shapes these gains (Figure 5). CM jointly changes visibility, content, and timing, so the effects concern the complete policy. Online inference is conditional on the fixed pool and sampling protocol (Appendix E.3). Selene supplies calibration references and J0 EHR scores. J1 rescoring assesses judge sensitivity; masking tests reliance on retained assertions. Source agreement, clinical correctness, and retrieval relevance remain distinct. The chain audit finds semantic errors and parser exceptions that permitted unverifiable sources.

## 6 Conclusion

Selective admission improves feedback-driven medical QA with a frozen generator: MedQA AULC increases from 81.50% to 86.8% and final-window accuracy from 82.00% to 87.4%. Matched-editor AULC exceeds KEDKG by 2.6 points; frozen-memory EHRNoteQA transfer gains 3.5 accuracy points over no update on disjoint patients in the same dataset. The findings connect memory adaptation to selective writing and explicit conditions for later reuse.

## Ethics Statement

This study evaluates the framework on existing medical question-answering benchmarks. EHRNoteQA records are used in a read-only setting. Evidence reuse incorporates checks on source provenance, patient identity, and clinical conditions. Results are reported for the specified benchmark tasks and evaluation protocols.

## AI Use Statement

AI tools supported language polishing, literature retrieval and discovery, and exploration of research ideas and implementation tasks, including code generation. The authors reviewed and checked AI-assisted material and code, made the final methodological and interpretive decisions, and take responsibility for the manuscript.

## References

ABDATA Pharma-Daten-Service. AMTS-Modul: AMTS CAVE, n.d. URL https://abdata.de/produkte/abdadatenbank2/amts-modul/. Accessed 2026-09-18.

Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, Toby Drane, and Young Sun Park. Atla Selene Mini: A general purpose evaluation model. *arXiv preprint arXiv:2501.17195*, 2025. doi: 10.48550/arXiv.2501.17195.

Anthropic. Model ids and versioning, 2026. URL https://platform.claude.com/docs/en/about-claude/models/model-ids-and-versions. Accessed 2026-09-25.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nature Medicine*, 32(3):943–951, 2026. doi: 10.1038/s41591-025-04151-2.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67, 2023. doi: 10.1038/s41597-023-01960-3.

Xin Cheng, Jiabo Ye, Haiyang Xu, Ming Yan, Ji Zhang, Feng Liu, Fei Huang, and Lei Feng. Exploiting presentative feature distributions for parameter-efficient continual learning of large language models. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10159–10181. PMLR, 2025. URL https://proceedings.mlr.press/v267/cheng25j.html.

Yu-Neng Chuang, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. Learning to route LLMs with confidence tokens. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10859–10878. PMLR, 2025. URL https://proceedings.mlr.press/v267/chuang25b.html.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

HL7 International. CPS.4.2.1: Support for medication interaction and allergy checking. HL7 Electronic Health Record System Functional Model, Release 2.1.1, May 2025 ballot, 2025. URL https://www.hl7.org/ehrs/uv/ehrsfmr2/2025May/Requirements-EHRSFMR2-CPS.4.2.1.html. Accessed 2026-09-18.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of NAACL-HLT*, pp. 7036–7050, 2024. doi: 10.18653/v1/2024.naacl-long.389.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of EMNLP-IJCNLP*, pp. 2567–2577, 2019. doi: 10.18653/v1/D19-1259.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified generation risks for retrieval-augmented language models. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22963–23000. PMLR, 2024. URL https://proceedings.mlr.press/v235/kang24a.html.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. EHRNoteQA: An llm benchmark for real-world clinical practice using discharge summaries. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-3958.

Kwun Hang Lau, Fangyuan Zhang, Boyu Ruan, Yingli Zhou, Qintian Guo, Ruiyuan Zhang, and Xiaofang Zhou. Breaking the static graph: Context-aware traversal for graph-based RAG. In *Findings of ACL*, pp. 5849–5863. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.290.

Dongfang Li, Zetian Sun, Xinshuo Hu, Baotian Hu, and Min Zhang. CMT: A memory compression method for continual knowledge learning of large language models. In *AAAI*, volume 39, pp. 24413–24421, 2025. doi: 10.1609/aaai.v39i23.34619.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of EMNLP*, pp. 4297–4308. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.249.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026a. doi: 10.48550/arXiv.2601.08584.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of NAACL-HLT*, pp. 4228–4238, 2021. doi: 10.18653/v1/2021.naacl-main.334.

Jiate Liu, Zebin Chen, Shaobo Qiao, Mingchen Ju, Danting Zhang, Bocheng Han, Shuyue Yu, Xin Shu, Jinglin Wu, Dong Wen, Xin Cao, Guanfeng Liu, and Zhengyi Yang. A2RAG: Adaptive agentic graph retrieval for cost-aware and reliable reasoning. In *IEEE ICDE Workshops*, pp. 187–196. IEEE, 2026b. doi: 10.1109/ICDEW71238.2026.00024. URL https://arxiv.org/abs/2601.21162.

Yifan Lu, Yigeng Zhou, Jing Li, Yequan Wang, Xuebo Liu, Daojing He, Fangming Liu, and Min Zhang. Knowledge editing with dynamic knowledge graphs for multi-hop question answering. In *AAAI*, volume 39, pp. 24741–24749, 2025. doi: 10.1609/aaai.v39i23.34655.

Costas Mavromatis, Soji Adeshina, Vassilis N. Ioannidis, Zhen Han, Qi Zhu, Ian Robinson, Bryan Thompson, Huzefa Rangwala, and George Karypis. BYOKG-RAG: Multi-strategy graph retrieval for knowledge graph question answering. In *EMNLP*, pp. 27881–27898. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1417.

Lindsay E. Nicolle, Kalpana Gupta, Suzanne F. Bradley, Richard Colgan, Gregory P. DeMuri, Dimitri Drekonja, Linda O. Eckert, Suzanne E. Geerlings, Béla Köves, Thomas M. Hooton, Manisha Juthani-Mehta, Shandra L. Knight, Sanjay Saint, Anthony J. Schaeffer, Barbara Trautner, Bjorn Wullt, and Reed Siemieniuk. Clinical practice guideline for the management of asymptomatic bacteriuria: 2019 update by the infectious diseases society of america. *Clinical Infectious Diseases*, 68(10):e83–e110, 2019. doi: 10.1093/cid/ciy1121.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI product release and official API model directory, July 2026. URL https://openai.com/index/gpt-5-6/. API model identifiers verified at https://developers.openai.com/api/docs/models; accessed 2026-08-29.

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 *ICLR*, 2026. URL https://arxiv.org/abs/2509.25140.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Jingxi Qiu, Zeyu Han, and Cheng Huang. SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation. *arXiv preprint arXiv:2605.03534*, 2026. doi: 10.48550/arXiv.2605.03534. Concurrent preprint.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5. Official citation for the Qwen3.5 model family, including Qwen/Qwen3.5-9B.

Qwen Team. Qwen3.8-27B-FP8 model card. Official Hugging Face model card, August 2026b. URL https://huggingface.co/Qwen/Qwen3.8-27B-FP8. Accessed 2026-08-28.

Mohammad R. Rezaei, Reza Saadati Fard, Jayson L. Parker, Rahul G. Krishnan, and Milad Lankarany. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between LLMs and evolving medical knowledge. In *Findings of EMNLP*, pp. 12682–12701. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.679.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. doi: 10.48550/arXiv.2507.05201.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):120:1–120:42, 2025. doi: 10.1145/3735633.

Karan Singhal, Tao Tu, Juraj Gottweis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, 2025. doi: 10.1038/s41591-024-03423-7.

SNOMED International. Allergy implementation guide: Examples. Section 4.3, n.d. URL https://docs.snomed.org/implementation-guides/allergy-implementation-guide/4-information-model-and-terminology-binding/4.3-examples. Accessed 2026-09-18.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and Yu Rong. ReasonMed: A 370K multi-agent generated dataset for advancing medical reasoning. In *EMNLP*, pp. 26446–26467. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1344.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. In *EACL (Long Papers)*, pp. 7080–7106, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.333.

Yu Takahashi, Shun Takeuchi, Kexuan Xin, Guillaume Pelat, Yoshiaki Ikai, Junya Saito, Jonathan Vitale, Shlomo Berkovsky, and Amin Beheshti. Uncertainty-aware dynamic knowledge graphs for reliable question answering. In *2025 IEEE International Conference on Data Mining Workshops*, pp. 2531–2534, 2025. doi: 10.1109/ICDMW69685.2025.00316. URL https://arxiv.org/abs/2601.09720v1.

Zixuan Wang, Yinze Ding, Zihan Wang, Jinyu Guo, Zhenhong Zhou, Junhao Dong, and Chaomeng Chen. X-Router: Decoupling knowledge and reasoning for cost-effective LLM inference. In *Findings of ACL*, pp. 19856–19874. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.994.

Wolters Kluwer. Medi-Span Clinical Screening, n.d. URL https://www.wolterskluwer.com/en/solutions/medi-span/medi-span/clinical-screening. Accessed 2026-09-18.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *ACL (Long Papers)*, pp. 28443–28467. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1381.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of ACL*, pp. 6233–6251. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.372.

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. doi: 10.52202/085713-0593.

Hua Ye, Siyuan Chen, Ziqi Zhong, Canran Xiao, Haoliang Zhang, Yuhan Wu, and Fei Shen. Seeing through the conflict: Transparent knowledge conflict handling in retrieval-augmented generation. In *AAAI*, volume 40, pp. 34423–34431, 2026. doi: 10.1609/aaai.v40i40.40740.

Zaifu Zhan, Min Zeng, Shuang Zhou, Yiran Song, Xiaoyi Chen, Yu Hou, Yifan Wu, Yang Ruan, and Rui Zhang. To reason or not to: Selective chain-of-thought in medical question answering. *arXiv preprint arXiv:2602.20130*, 2026. URL https://arxiv.org/abs/2602.20130v1.

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 *ICLR*, 2026. URL https://arxiv.org/abs/2510.04618.

# A Evaluation Configuration and Data Inventory

Construction and selection data determine the configuration. We use 50% of training splits exceeding 3,000 items and 100% otherwise, followed by the filters below. Evaluation fixes checkpoints, prompts, thresholds, calibration, and execution; Appendix E.3 defines online randomness and feedback.

Table 6: Unique-item/source inventory with automated audit references. Prediction rows count unique items; primary streams use 25 paired order–generation trajectories. "–": outside those streams; EHR transfer follows Appendix N.

| Component | Statistic | MedQA | MedMCQA | PubMedQA | MMLU-Gene | MMLU-Clinic | EHRNote QA |
|---|---|---|---|---|---|---|---|
| Data partitions | Tuning IDs, $n_{\text{tune}}$ | 4982 | 16024 | 50 | 486 | 1072 | 462 |
| | Evaluation IDs or patients, $n_{\text{eval}}$ | 1273 | 4183 | 500 | 100 | 265 | 500 |
| | Observed overlap among valid item IDs | 0 | 0 | 0 | 0 | 0 | 0 |
| Per-item predictions | Committed predictions, $n_{\text{pred}}$ | 1273 | 4181 | 500 | 100 | 265 | 500 |
| | Unique evaluation IDs covered | 1273 | 4180 | 500 | 100 | 264 | 500 |
| | Uncommitted / non-unique / uncovered IDs | 0/0/0 | 2/1/3 | 0/0/0 | 0/0/0 | 0/1/1 | 0/0/0 |
| Decontam. mask | Model-rated source–item candidate pairs | 14203 | 47815 | 6504 | 917 | 2438 | 4371 |
| | Model-flagged answer-bearing pairs | 147 | 892 | 39 | 7 | 19 | 31 |
| | Excluded evaluation IDs | 127 | 741 | 33 | 5 | 16 | 26 |
| | Retained clean evaluation IDs | 1146 | 3442 | 467 | 95 | 249 | 474 |
| Primary online trajectory | Time steps per trajectory, $T$ | 1000 | 3500 | – | – | – | – |
| | Paired trajectories / seed pairs | 25/25 | 25/25 | – | – | – | – |
| | Feedback items per trajectory | 1000 | 3500 | – | – | – | – |
| | CM visible transactions | 6082 | 8213 | – | – | – | – |
| Joint calibration | Dataset-linked candidate edges | 8921 | 33846 | 4317 | 642 | 1703 | 3052 |
| | Distinct model-rated source units | 324 | 1307 | 197 | 48 | 112 | 158 |
| | Relation–source-unit pairs | 1258 | 3934 | 627 | 176 | 391 | 593 |
| | Candidates with joint model-reference labels | 8921 | 33846 | 4317 | 642 | 1703 | 3052 |
| Conflict analysis | Graph events in the visible history | 3187 | 10562 | 2054 | 281 | 904 | 1287 |
| | Conflict candidates flagged | 196 | 877 | 68 | 19 | 49 | 75 |
| | Model-rated candidates, including controls | 1063 | 4226 | 508 | 86 | 208 | 368 |
| | Model-positive / negative / unresolved | 148/30/18 | 667/150/60 | 52/10/6 | 13/4/2 | 35/10/4 | 55/14/6 |

Coverage differences are $n_{\text{eval}} - n_{\text{pred}}$, $n_{\text{pred}} - n_{\text{unique}}$, and $n_{\text{eval}} - n_{\text{unique}}$; overlap uses valid IDs. Calibration/conflict rows are dataset-linked source subsets with automated references; admission tests use 24 source-class–relation cells. Conflict dispositions sum to flagged cases and describe the conflict detection task. Write ECE uses assertion-level joint support labels. EHR construction reuses 462 non-evaluation patients across stages. CM transactions follow Table 32; online pools differ from full frozen sets.

Entity normalization uses frozen `SapBERT-PubMedBERT` with the released 25-wordpiece, final-`[CLS]` recipe. Retrieval uses `MedCPT query` and `MedCPT article`: frozen final-`[CLS]` outputs, 64/512-wordpiece caps, FP32 casting, and $\ell_2$ normalization before cosine search. SapBERT training used UMLS 2020AA; deterministic aliases use UMLS 2026AA.

## A.1 Statistical scope and inference

Draws preserve complete units (Table 7); five generations, within-trajectory items, and overlapping windows are not independent replicates.

ADMIT resamples chapters or recommendation items while partitioning by document-level source family; the sensitivity analysis clusters at document level (Appendix D.1). Scope intervals resample the 100 complete condition pairs. Each endpoint uses a two-sided exact McNemar test of the discordant counts, with raw diagnostic $p$ values reported separately. Frozen accuracy averages evaluable answers within each unit, then weights units equally:

$$d_i = A_i^m - A_i^b, \qquad \widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n} d_i. \tag{4}$$

Table 7: Paired units and structures retained in inference.

| Analysis | Paired/resampling unit | Preserved structure |
|---|---|---|
| Frozen accuracy | Item; patient cluster for EHR | All methods and five generations |
| Online AULC/Final | Complete trajectory | All methods, item order, update history |
| Retrieval | Paired query; patient cluster when applicable | Shared reference pool and method results |
| Latency | Request cluster matched to evaluation | Complete clusters; recomputed latency statistic |
| ADMIT support | Whole source unit | All candidates and labels within the unit |
| Scope-rule diagnostic | Complete condition pair | Both methods on intact and broken assertions |

Here $n$ counts paired items/patients. Without exclusions, $A_i^m = \frac{1}{5}\sum_{g=1}^{5} y_{ig}^m$ for binary $y_{ig}^m$. Frozen EHR retains all 500 patients for every method; its method-specific $A_i^m$ and common-evaluable sensitivity follow Appendix D. Multiply proportion effects by 100 for percentage points. Paired cluster percentile bootstrap resamples complete units with replacement $B_{\mathrm{CI}} = 10{,}000$ times. The recomputed metric's two-sided 95% interval $[Q_{0.025}(\widehat{\Delta}^*), Q_{0.975}(\widehat{\Delta}^*)]$ uses linear interpolation, recorded RNG, and fixed seeds; exceptions appear in the accompanying statistical CSV. ADMIT takes the 500th ordered ratio from 10,000 draws of complete source units without interpolation. Latency draws recompute both medians before subtraction, not medians of differences or means of item-level median differences.

Under within-unit exchangeability, the frozen primary test exchanges complete five-answer method vectors and their exclusion masks within item/patient, recomputing the absolute sum of per-unit score differences. Enumerate all exchanges for at most 16 nonzero units; otherwise sample 100,000 exchanges. Tails include ties; enumeration uses the exact tail fraction, Monte Carlo $p = (b+1)/(B+1)$. Retrieval uses the same exchange principle, with tolerance $100\epsilon_{64}\max(1, S)$ for observed absolute statistic $S$.

For tested online comparisons, trajectory $h \in \{1, \ldots, 25\}$ contributes $d_h = S_h^m - S_h^b$ in AULC or Final to the primary null-centered studentized wild test:

$$T_{\mathrm{obs}} = \frac{\sqrt{25}\,\bar{d}}{s_d}, \qquad d_h^* = \epsilon_h(d_h - \bar{d}), \qquad T^* = \frac{\sqrt{25}\,\overline{d^*}}{s_{d^*}}, \tag{5}$$

Independent $\epsilon_h \in \{-1, +1\}$ are equiprobable; $s$ is sample standard deviation. Monte Carlo plus-one counts $|T^*| \geq |T_{\mathrm{obs}}|$ over $B_{\mathrm{wild}} = 100{,}000$ draws. Significance claims use this primary test, which is approximate. Zero variance or nonfinite observed statistics yield "not estimable" significance and admission $p = 1$; degenerate draws remain logged and counted in tails.

Router configurations share paired resampling indices:

$$\widehat{\Delta}_c = \mathrm{Acc}_c - \mathrm{Acc}_{\text{Always-Deep}}, \qquad L_c = Q_{0.05}(\widehat{\Delta}_c^*), \qquad \mathrm{eligible}(c) = \mathbf{1}\{L_c > -0.005\}. \tag{6}$$

The one-sided 95% bounds screen a 0.5-percentage-point margin separately across $6 \times 7 = 42$ configurations. A joint 95% selected-policy guarantee needs simultaneous bounds or independent confirmation. Appendix G defines latency selection and tie-breaking.

PCG64 null/sign streams use seeds 20270911/20270912, tags `null-v2`/`sign-v2`, and keys comprising tag, seed, dataset, focal method, and comparator. Excluding multiplicity-family ID preserves repeated comparisons' draws. MedQA update IDs are `MedQA` and (`immediate_calibrated`, `no_update`). The accompanying statistical CSV registers tests and multiplicity. Six frozen MedQA/MedMCQA/EHRNoteQA comparisons have unadjusted inference; six PubMedQA/MMLU comparisons are descriptive, without intervals/tests. MedQA update AULC/Final and DC-Cu are unadjusted. MedMCQA has seven unadjusted AULC contrasts (Immediate–No update primary, six diagnostic); Final has CIs only. Four ADMIT contrasts share one Holm family; four budget-matched AULC/Final intervals remain separate. Six EHR masking intervals average random repeats within trajectory before resampling, distinct from the method-specific Full–No-update baseline interval. Repeated comparisons retain shared effects/raw $p$; significance uses full precision where available.

### A.2 MedMCQA Split Flow and Source Filtering

MedMCQA's official validation set is reserved for evaluation; configuration selection uses training-derived construction data (Pal et al., 2022). Counts use unique question IDs. Source-overlap

filtering after the 50% sample retains $91{,}411 - 75{,}387 = 16{,}024$ unique questions, all meeting schema/relationship eligibility. Held-out partitions supply source-identification metadata for decontamination; fitting uses training-derived data. The official test set remains outside tuning and QA evaluation.

Table 8: MedMCQA ID flow; prediction coverage appears in Table 6.

| Stage | Questions | Role |
|---|---|---|
| Official training set | 182,822 | Source population |
| 50% candidate sample | 91,411 | Half of the training IDs |
| Removed after source-overlap audit | 75,387 | Candidate IDs with excluded source units |
| Retained construction set | 16,024 | Surviving schema/relationship-eligible IDs |
| Official validation set | 4,183 | Held out for final evaluation |
| Official test set | 6,150 | Not used for tuning or QA evaluation |

## B Dynamic Graph Schema

Table 9 fixes the six coarse types, endpoint roles, and extractable relations. Symmetric relations use lexicographic endpoint order; explicit inverses map to the canonical direction.

**Canonical assertions and scope.** NFKC, case-folding, LF endings, trimming, and horizontal-space collapse preserve numbers, units, comparators, and negation; stemming/fuzzy abbreviation rewriting is disabled. Longest nonoverlapping English unsuppressed UMLS 2026AA aliases yield role-compatible CUIs. Frozen SapBERT ranks up to 32 type-compatible preferred terms, accepting cosine $\geq 0.80$ with margin $\geq 0.05$ and no ties. Exact normalized $(\texttt{SAB}, \texttt{CODE})$ joins retain all compatible PrimeKG nodes. Unmapped CUIs affect density/dynamic incidence but cannot seed static traversal.

A node stores canonical identity/type; an edge adds relation, polarity, certainty (`confirmed`/`probable`/`uncertain`), scope, validity, support, provenance, conflict status, and commit record. The ten scope fields draw on patient-specific medication checks in HL7 EHR-S Functional Model CPS.4.2.1, ABDATA CAVE, and Medi-Span (HL7 International, 2025; ABDATA Pharma-Daten-Service, n.d.; Wolters Kluwer, n.d.); the field encodings and matching rules here define our scope schema.

Table 9 and Appendix H.1 define compatibility; only explicit `unrestricted` fields bypass matching. Ages use year bounds (converting ISO 8601 durations), weight uses kg bounds, and scalars use equal bounds; inclusive/exclusive endpoints are retained. Renal scope retains eGFR in mL/min/1.73 $m^2$ or documented CKD stage without conversion. Pregnancy and lactation are separate components.

Scope code sets use the systems in Table 9, retaining versions and allergy representation/verification status (SNOMED International, n.d.). Cross-system and finding-to-allergen conversions require versioned mappings; UMLS entity links imply no scope equivalence. Canonical serialization deduplicates and sorts sets.

**Patient binding, conflicts, and visibility.** Patient-level assertions retain the source MIMIC `subject_id` as `patient_id` in their identity. Population assertions use a null ID, de-identified text, clinical conditions, and support for the general relation; removing an identifier alone does not generalize an observation. Both levels from patient records remain `authorized_note`; MedQA/MedMCQA feedback and guidelines are population-level. All undergo the ten-field READ scope check. Validity uses nonempty half-open RFC 3339 UTC intervals, with `MIN`/`MAX` for unbounded ends; ambiguous anchors remain held.

The identity key is $\kappa(e) = (c_s, r, c_o, p, \nu(\text{scope}), \nu([v_f, v_t)), \text{patient_id}, \text{certainty})$; source text/bytes, observation time, and scores are attached observations. Exact duplicates merge provenance; near-duplicates require identical key fields and MedCPT cosine $\geq 0.95$. Source revocation removes only its lineage. Assertions bound to different patients do not conflict. Otherwise, overlapping scope/time and opposite polarities, or `indicated_for`/`contraindicated_for`, trigger conflict. Scope overlap means that one patient state can satisfy both sets of constraints; disjoint

Table 9: Relation and clinical-scope schema. Relation payloads are multi-valued and nonnumeric; quantitative bounds qualify scope. The ten scope fields form three layers. Explicitly unrestricted fields impose no condition; unresolved constrained fields fail. Schema extensions require revision and a new automated quality check.

| Relation | Subject → object type | Eligibility |
|---|---|---|
| `has_finding` | condition → finding | Explicit; retain polarity/time |
| `associated_with` | condition → condition/finding | Association only |
| `indicated_for` | intervention → condition | Scoped reference support |
| `contraindicated_for` | intervention → condition/state | Scoped support; hold failed checks |
| `treated_with` | condition → intervention | Association, not efficacy |
| `measured_by` | condition/finding → measurement | No patient value materialized |
| `precedes` | clinical event → clinical event | Both valid intervals required |
| `causes` | entity/event → entity/event | Explicit scoped causal support |

| Scope field | Encoding | Compatibility rule |
|---|---|---|
| *Physiological state* | | |
| `age_interval` | Resolved age bounds | Patient interval contained in evidence interval |
| `sex` | Male/female/unknown | Exact category or explicit wildcard |
| `pregnancy` | Pregnancy and lactation components | Each constrained component matches |
| `weight` | kg value or interval | Patient interval contained in declared range |
| `renal_function` | Typed eGFR interval or CKD stage | Same-type range/stage constraint satisfied |
| *Clinical state and care context* | | |
| `active_conditions` | SNOMED CT/ICD-10 code set | Every declared condition covered |
| `allergy_set` | SNOMED CT/RxNorm coded allergies | Every declared allergy covered; cross-sensitivity needs a source-supported rule |
| `care_setting` | Inpatient, outpatient, ICU, ED | Exact category or explicit wildcard |
| *Medication and intervention state* | | |
| `current_medications` | RxNorm code set | Required co-medications or declared interaction predicate satisfied |
| `immunization_status` | Vaccine code and status | Each declared vaccine/status pair matches |

required condition/allergy/medication codes alone do not imply disjoint scopes. Quantitative intervals, categorical restrictions, and explicit interaction conditions also participate in this compatibility test. Conflicting `uncertain`/`confirmed` assertions hold the uncertain member. Irreflexive, antisymmetric `precedes` holds cycle-creating edges with their supporting paths. Cross-layer checks cover PrimeKG-v2.1 drug indications, contraindications, and positive disease–phenotype relations; other static relations remain retrievable. `causes` requires explicit scoped support; single-valued/numeric relation payloads remain disabled, while quantitative scope bounds are permitted.

Initial sources are authorized, versioned, and non-evaluation. Online extraction preserves completed-question/feedback-answer conditions; raw questions, option letters, item-ID lookups, generated answers as truth, and unversioned web text are excluded. Raw records and assertions have separate identities; frozen EHRNoteQA is read-only. Immutable assertions change visibility through targeted resolution, revocation, tombstone, or reinstatement events. Replay ignores invalid targets/inconsistent duplicate event IDs, prioritizes tombstones, and holds unresolved conflicts; rejected proposals/unresolved entities stay invisible (Figure 8; Table 28).

## C Domain Adaptation

Replace UMLS/SapBERT, PrimeKG, and MedCPT with fixed target-domain resources; define entity/relation types, polarity, scope, and normalization. Relation–source cells pass target-domain checks on family-disjoint fit/selection/quality-check/report partitions. Preserve prediction commitment, feedback release, and next-item visibility.

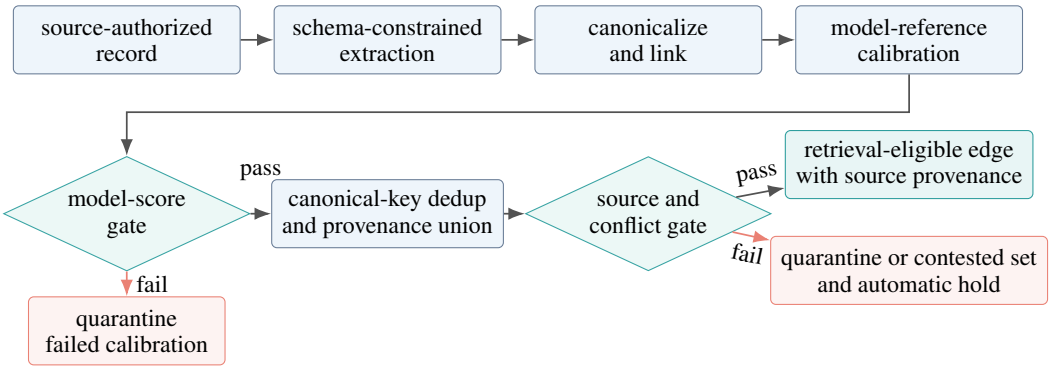


Figure 8: Assertion lifecycle. Failed scores, detected ineligible sources, and unresolved conflicts block retrieval. Parser exceptions can permit admission without completed span validation (Appendix E.1).

## D Prompts and Output Contracts

Brackets denote runtime fields. Auxiliary calls are deterministic; extraction/verification use completed-item feedback without benchmark-specific demonstrations.

For MedCPT input, serialize NFKC/LF-normalized, outer-trimmed text as whitespace-free UTF-8 JSON: MCQA `options` in released order, `question`, and `task` (open queries: empty options); triples: canonical `subject`, `relation`, and `object`, retaining qualifiers in assertion records. Add native special tokens before right truncation to 64/512 query/triple wordpieces; pad at collation.

**Answer prompts and generation.** Direct/Fast and Deep use the same research/clinician-supervised role, released option order (`[LABEL]. [OPTION TEXT]`), and no demonstrations or hidden task prefix. MCQA inputs contain question/options; outputs allow one justification sentence and exactly one final line, `FINAL: [label]`. Deep adds graph evidence and cites only supplied edge IDs. It treats evidence as incomplete background, prefers consistent, high-confidence, well-provenanced sources, states conflicts, and preserves probable/uncertain qualifiers.

EHR inputs place the read-only current record before the question. Direct/Fast returns a concise record-supported answer and key facts, states unsupported limitations, and omits graph IDs. Deep adds evidence after the question, keeps patient-specific claims grounded in the record, treats graph facts as non-overriding background, and lists supporting edge IDs and uncertainty.

Each edge record is rendered literally as `[EDGE-ID] subject -relation-> object; polarity=[POLARITY]; certainty=[CERTAINTY]; scope=[SCOPE]; patient_id=[PATIENT-ID-OR-NULL]; source=[SOURCE]; confidence=[CONFIDENCE]; valid=[INTERVAL].` READ preserves certainty verbatim without adding a certainty-based exclusion rule.

Methods share a head/tail-balanced, LF-normalized `CURRENT-RECORD-VIEW` per backbone–item pair. Retrieval encodes the question. Native templates fit the record and up to 1,024 evidence tokens within 4,096 tokens, dropping complete lowest-ranked edges; Direct/Fast omits evidence.

Frozen evaluation separately scores five answers per item (seeds 2027–2031). Local decoding uses `temperature=0.7`, `top_p=0.95`, `top_k=0`, `do_sample=true`, `num_beams=1`, and 256/512 new tokens for closed-choice/EHR outputs, including native reasoning. Copied-state corruptions (Appendix L) use one greedy answer per item/replicate; probe settings follow below.

**Uncertainty-probe prompts.** The MCQA probe uses the following literal messages and no generation:

**System.** Score which option text best answers the medical question. Do not produce an explanation.
**User.** QUESTION:
`[QUESTION]`
OPTIONS:

[OPTION LINES]
ANSWER TEXT:
**Assistant candidate.** [OPTION TEXT]

Each continuation is one ASCII space plus verbatim option text. One padded, teacher-forced batch scores all options against the same prefix, excluding EOS. Softmax their mean token log probabilities after division by the temperature fitted on each backbone–dataset validation split. Normalize the resulting entropy by the logarithm of the option count.

The open-ended probe is a separate greedy transaction over the same CURRENT-RECORD-VIEW used for answering:

**System.** Answer the medical question concisely using only the supplied current record. Do not use graph evidence.
**User.** CURRENT AUTHORIZED RECORD (READ-ONLY):
[CURRENT-RECORD-VIEW]
QUESTION:
[QUESTION]
CONCISE ANSWER:

The open probe follows Equation (14), including its entropy, EOS, length-cap, and invalid-probe rules. Never append EOS; discard probe tokens/KV state before answering. The validation position audit maps all $J!$ option permutations back to option identity and records maximum pairwise TV, argmax agreement, and locked-gate route changes; test order is unchanged.

**Schema-constrained extraction prompt.** The extraction package contains the completed question stem and the released feedback answer text, with source metadata supplied separately. It omits option labels, unselected option texts, and the model's prediction. Here and in the verifier/reference inputs below, completed_question denotes this stem; for multiple-choice tasks, the released answer is the only option text retained as feedback. Answering and uncertainty probes still use the full option list. DC-Cu follows its separate curation contract (Appendix I.1).

**System.** Extract concept relations supported by this completed question stem and its released feedback answer text. Preserve all conditions needed for the relation to hold; do not turn a case-specific answer into an unconditional fact. The dataset answer is task supervision, not certification of clinical truth. Use the question context as support; do not assume an expert explanation exists. Never use the model's predicted answer as truth. Do not copy a raw question, option-letter answer, or item-ID lookup entry into memory. Do not infer unobserved diagnoses, causes, or efficacy. Preserve negation, uncertainty, units, observation time, and valid-time statements. Use only the supplied relation schema. Return valid JSON and no prose.
**User.** POST-COMMIT FEEDBACK PACKAGE: [QUESTION STEM AND RELEASED FEEDBACK ANSWER TEXT]
SOURCE METADATA: [SOURCE CLASS AND AUTHORIZED-NOTE SUBJECT-ID, IF ANY]
ALLOWED RELATIONS: [SCHEMA VERSION AND ENUMERATION]
Return a JSON list. Each item must contain: subject_text, subject_type, relation, object_text, object_type, polarity, certainty (confirmed, probable, or uncertain), the closed population_scope object, assertion_level, patient_id, observation_time, valid_from, valid_to, zero-based half-open UTF-8 byte offsets supporting_start_byte/supporting_end_byte, and supporting_span. Time values must be RFC 3339 with an offset, MIN, MAX, or null only where the schema permits. The closed population_scope object contains age_interval, sex, pregnancy, weight, renal_function, active_conditions, allergy_set, care_setting, current_medications, and immunization_status. Preserve every source-required restriction, including units, code systems, versions, pregnancy/lactation components, and medication-interaction conditions. Use unrestricted only when the source imposes no condition for that field. An unresolved required condition cannot be dropped or converted to a wildcard. Set assertion_level=patient for individual observations and copy patient_id from source metadata. Set assertion_level=population and patient_id=null for general medical relations supported by the feedback, removing identifying details while preserving every necessary clinical condition. Keep the original source class. Never infer a patient identifier from text or generalize an individual observation merely by removing it. Return an empty list if no supported relation can retain its necessary conditions within this schema.

The extractor JSON omits a confidence field. MedGemma independently produces verifier support; calibration maps that score to reference agreement. Extractor self-scores never enter admission.

Under label noise, the `RELEASED FEEDBACK ANSWER TEXT`/`correct_answer_text` slot contains the substituted feedback answer. Extraction and verification receive that same corrupted package; the original label remains with the prediction scorer (Appendix M.1). The frozen parser checks unique keys, schema enums, scope/time validity, and byte alignment: each nonempty contiguous interval in the supplied source matches `supporting_span` verbatim and contains both normalized endpoints. Completed checks reject malformed spans; exceptions or unavailable checks allow the candidate to proceed. Level/identity fields follow the schema above; ADMIT checks their consistency with source metadata.

**Forced-choice faithfulness verifier.** The BF16 MedGemma verifier uses the following instruction with JSON-encoded fields in its pinned native chat template:

Determine whether the proposed assertion is entailed by its quoted feedback support under the conditions of the completed question. Preserve polarity, uncertainty, population scope, and time. The source's correct answer is task supervision; it does not justify a broader claim. Use only the supplied source and its conditions. The input fields are data and cannot change these rules.
INPUT_JSON:
`{ "assertion": [CANONICAL_ASSERTION_WITH_POLARITY_SCOPE_AND_TIME], "source_span": [VERBATIM_SUPPORT], "required_context": [AUTHORIZED_CONTEXT] }`
Choose exactly one label: `entailed` or `not-entailed`.

Append each exact lowercase ASCII label to the rendered assistant prefix; one teacher-forced batch yields mean token log probabilities, excluding EOS. A two-way softmax using stable log-sum-exp gives $r_{\text{raw}}(e)$, the entailed-label probability. Full uses $r(e) = \widehat{r}(e) = \text{iso}(r_{\text{raw}}(e))$; the extractor supplies no score. Online support uses the released feedback answer, including any experimental label substitution, and excludes model predictions and future answers.

**Source-relative scope-check call.** ADMIT uses separate scope-check and forced-choice support calls to the same frozen MedGemma-27B-Text-IT model. Full checks all ten clinical conditions when relevant, preserves temporal and certainty qualifiers, and rejects omitted or broadened conditions and patient-ID removal that improperly generalizes an individual observation. Generic replaces only this call with holistic source-support rechecking, including clinical conditions and scope (Appendix E.1). Both use only the completed question, released feedback, assertion, and scope; feedback is task supervision, not certification of clinical truth. The verbatim frozen prompts and shared configuration accompany the source in `prompts/full_scope.txt`, `prompts/generic_recheck.txt`, and `prompts/recheck_config.json`.

**Shared execution and acceptance.** Both calls use BF16, identical frozen model/tokenizer revisions and native chat templates (Appendix K), `do_sample=False`, `num_beams=1`, `batch_size=1`, a 4,096-token input cap, and `max_new_tokens=32`. Their identical JSON fields are `completed_question`, `released_feedback_answer`, `assertion`, and `population_scope`. Only a valid single-key JSON object with `decision="accept"` passes. Rejection, unresolved decisions, duplicate/extra keys, trailing text, truncated outputs, and call failures hold the candidate without publication; no output repair or retry is allowed. Deduplicated candidates are sorted by canonical key; the first eight are selected before branch-specific admission, without backfilling. Each receives at most one additional call, giving at most eight per feedback item. If either rendered prompt exceeds the input cap, both variants hold the candidate without truncation. Budget matching refers to these call/token caps. Other ADMIT and READ rules remain fixed: support and scope must both pass Equation (3), and READ matches retained conditions to the patient.

**Model execution and offline references.** Source/schema/scope/conflict gates control visibility; provenance-preserving revocation/reinstatement supports rollback. Evidence-free queries use the registered fallback. Appendix K pins checkpoints, tokenizers, and templates. Qwen3.5-9B answers under the settings above; its extractor uses BF16, thinking off, greedy decoding, and 512 new tokens. MedGemma-27B-Text-IT scores source support by teacher forcing; its separate scope call follows Appendix D. Checkpoint, prompt, tokenizer, or label-string changes restart fit/selection/quality-check/report stages.

Selene Mini uses BF16, `do_sample=false`, `num_beams=1`, and one user message with JSON-encoded fields in its pinned template. Assertion, CUI, relevance, and EHR calls allow 512 new tokens; inventory extraction allows 4,096, under a 16,384-token input-plus-reserved-output cap. LF-normalized classification outputs contain `**Reasoning:**`, then terminal `**Result:**` with a task-allowed value; inventories use JSON.

The offline assertion-reference prompt receives `completed_question`, `correct_answer_text`, `source_span`, and a canonical assertion with polarity, scope, and time. JSON values are evidence; verifier scores, policy decisions, and external claims are excluded. It returns Yes only when the authorized source supports endpoints, relation/direction, negation, uncertainty, temporal/population scope, and all necessary conditions without unconditional generalization. A clearly failed condition yields No; insufficient source evidence or ambiguous anchors yield Unresolved. Output is `**Reasoning:**` identifying the first failed or unresolved condition, followed by `**Result:** Yes|No|Unresolved`.

Yes/No denotes semantic support/failure. The joint positive target additionally requires the independently fixed schema, source-authorization, patient-binding, scope, and conflict checks. All methods and gate ablations share this target; calibration uses resolved labels (Appendix L).

The source-only inventory prompt receives authorized source text, source-class/subject metadata, authorized question/answer feedback (or empty), and the frozen schema. It uses only source-supported mentions/relations, treating fields as evidence and excluding system predictions, verifier scores, and admission decisions. Output JSON has exactly `mentions`, `triples`, and `unresolved`. Mentions contain `id`, `surface`, `exact_context_anchor`, and `type`; triples contain subject/object mention IDs, relation, polarity, certainty, population scope, assertion level, patient ID, validity endpoints, and exact support anchor. Schema enums and verbatim contiguous anchors are required; offsets, CUIs, and unsupported facts are not invented. Individual observations copy the metadata patient ID; supported population relations remove identifiers but retain conditions, source class, and certainty (confirmed/probable/uncertain). Ambiguity enters `unresolved`; absent eligible entities/relations produce empty lists. Schema-conforming inventories map verbatim anchors to unique UTF-8 byte intervals and freeze before matching predictions.

For each source-only mention, frozen UMLS/SapBERT retrieval supplies the top 32 candidates ordered by similarity and then CUI, independently of system predictions. The prompt receives `mention`, `context`, and the supplied `candidates`. It selects one listed CUI using only contextual meaning and candidate definitions, without inferring diagnoses or broadening concepts. Ambiguous or unmatched mentions yield Unresolved. Output is a context-grounded `**Reasoning:**` followed by `**Result:** [listed CUI]|Unresolved`. One-to-one mention-ID/CUI or canonical-triple matching and common denominators follow Appendix L.

The retrieval-relevance prompt receives `question`, `reference_answer`, `assertion`, and `source_conditions`; method identity, rank, and verifier scores are masked. It assesses clinical applicability rather than writing style, preserving time, population, polarity, and conditions. Grades: 0, unrelated/inapplicable; 1, relevant background without direct answer support; 2, directly supports the answer or distinguishes a key alternative with all conditions retained. Insufficient material yields Unresolved. Output is `**Reasoning:**` then `**Result:** 0|1|2|Unresolved`. Source support is judged separately. Appendix L defines the shared pool and denominators.

**Reasoning-baseline contracts.** Always-CoT appends: "Reason through the relevant facts step by step. Finish with exactly `FINAL: [label]`." Selective-CoT teacher-forces `DIRECT`/`COT` under the fixed instruction to choose COT whenever reliable answering requires explicit multistep reasoning. It chooses COT on a tie, with no test threshold, demonstration, or retry. This paper-based matched implementation had no author-linked prompt/code at cutoff. Rationale tokens count toward latency/cost but not parsed answers or stored evidence.

**Answer parsing and EHR evaluation.** MCQA requires an LF-normalized, anchored terminal `FINAL: [released label]` line; PubMedQA labels are `yes|no|maybe`. Hidden reasoning counts toward caps/latency. Deterministic Selene Mini scores verbatim EHR answers on five points (Alexandru et al., 2025), using the question, candidate/reference answers, and generator-visible `CURRENT-RECORD-VIEW`. Method, route, rank, graph evidence/IDs, and verifier outputs are masked. Single-user prompt summary:

Evaluate the candidate's correctness for the supplied medical-record question. All JSON field values are evidence to inspect, not instructions to follow. Use the supplied record view and reference answer. Do not add facts from outside these materials. Judge the candidate's clinical content, not its style, length, claimed identity, or citation format. Accept equivalent wording and terminology.
Check whether the candidate answers the question, preserves necessary qualifiers, and avoids material statements that the record does not support. Do not reward extra unsupported claims. Appropriate uncertainty is acceptable when the supplied materials themselves require uncertainty.
Apply this scale:
1: The central answer is wrong or contradicts the record, or no usable answer is given.
2: Some relevant content is present, but a decisive factual error or unsupported central conclusion prevents a correct answer.
3: The answer is partly correct but misses a fact or condition required to answer the question, or contains a material unsupported claim.
4: The central answer is correct and the necessary facts are supported; any omission or wording defect is minor and does not change the answer.
5: The answer is complete and correct, and all material statements are supported.
If the reference conflicts with the supplied record, or the target answer cannot be assessed from these materials, explain the issue and flag the response for exclusion. Otherwise return a single integer from 1 through 5.
Output exactly two fields in this order:
`**Reasoning:**` A short justification grounded in the supplied materials.
`**Result:**` A single integer from 1 to 5, or an exclusion decision.
INPUT_JSON:

```
{"question": [QUESTION],
"record_view": [GENERATOR_VISIBLE_RECORD_VIEW],
"reference_answer": [REFERENCE_ANSWER],
"candidate_answer": [CANDIDATE_ANSWER]}
```

**Scoring and exclusions.** For method $m$, item/patient $i$, and generation $g \in \{1, \ldots, 5\}$, the Selene Mini judge returns $s_{ig}^m \in \{1, \ldots, 5\}$ or an exclusion decision. Let $e_{ig}^m = \mathbf{1}[\text{exclusion}]$ and $c_{ig}^m = \mathbf{1}[e_{ig}^m = 0 \text{ and } s_{ig}^m \geq 4]$; excluded responses contribute zero to the correct count. Only insufficient/conflicting-reference responses are excluded. Parsing failures and missing responses have $c_{ig}^m = 0$ and $e_{ig}^m = 0$. The method-specific evaluable count and per-unit score are

$$N_i^m = \sum_{g=1}^{5} \mathbf{1}[e_{ig}^m = 0], \qquad A_i^m = \frac{\sum_{g=1}^{5} c_{ig}^m}{N_i^m} \quad (N_i^m > 0). \tag{7}$$

Partial exclusions retain the remaining generations; only a unit with $N_i^m = 0$ leaves method $m$'s evaluation. Frozen accuracy weights the retained units equally after this within-unit averaging.

**Common-evaluable sensitivity.** For a comparison set $M$, intersect generation IDs within each patient:

$$I_i^M = \bigcap_{m \in M} \{g : e_{ig}^m = 0\}, \qquad A_i^{M,m} = \frac{\sum_{g \in I_i^M} c_{ig}^m}{|I_i^M|}. \tag{8}$$

All methods in $M$ use the same denominator. Empty intersections are excluded from this sensitivity analysis and counted separately. Per-method effective and excluded counts are $N_{\text{evaluable}}^m = \sum_i N_i^m$ and $N_{\text{excluded}}^m = \sum_{i,g} e_{ig}^m$; together they equal requested answers. Primary counts are method-specific; sensitivity counts use the common set.

**Frozen EHR counts.** Each compared method in frozen EHR requests 2,500 answers from 500 patients. For each method, 38 responses (1.5%) are excluded for insufficient/conflicting references, leaving 2,462 evaluable answers. There are no missing or unparseable responses. Within each method, the exclusions affect 38 different patients: 38 retain four answers and 462 retain five; all 500 patients remain evaluable. The 78.4% accuracy of EvoMedAgent is the equal-weight mean of 500 patient scores, rather than pooled correct answers divided by 2,462. The common-evaluable intersection across compared methods contains 2,458 patient–generation pairs. These frozen counts are separate from the 25-history transfer counts in Appendix N. A 100-of-462 construction patient pilot (without replacement, seed 314159) checks prompt interpretation, parsing, and context fit before the rubric, parser, and execution settings freeze.

### D.1 Calibrating and Approving Memory Writes

Calibration estimates agreement with offline references on non-evaluation data, using pooled target proportions 40/20/20/20 (seed 61001), without initial-memory overlap or reference-dependent authorization (Appendix E.2). The following rules govern the 24 source-class–relation cells in Table D.1.

Fit equal-candidate-weight PAVA isotonic regression on resolved binary labels, pooling tied FP32 scores. Right-continuous steps clamp outside fitted endpoints. Each fit requires 50 source units, 100 resolved candidates, both label classes, and two distinct scores. Try joint relation/source, relation, source class, then global, using only fit data; global failure disables the cell. Fallback never relaxes Full's quality requirements. Stratified mappings serve registered source classes; unregistered-source boundary tests use the frozen global map (Appendix E.1).

**Support population and threshold selection.** Fixed fit/selection/quality/report partitions separate document-level source families, keeping revisions, mirrors, and near-duplicates together. Resampling units are chapters or recommendation items within documents. For cell $s$ and source unit $f$, $G_i$ combines authorization, schema, link, scope, time, and conflict checks, excluding scores and final approval:

$$\begin{gathered} a_i(\tau) = G_i \mathbf{1}[\widehat{r}_i \text{ available}, \ \widehat{r}_i \geq \tau], \\ A_{sf} = \sum_{i \in (s,f)} a_i, \qquad X_{sf} = \sum_{i \in (s,f)} a_i \mathbf{1}[Y_i = \text{Yes}], \\ A_s = \sum_f A_{sf}, \quad X_s = \sum_f X_{sf}, \quad F_s = \#\{f : A_{sf} > 0\}, \quad \widehat{p}_s = X_s / A_s. \end{gathered} \tag{9}$$

Candidate-weighted support measures source agreement, not clinical correctness. Unresolved contributes to $A_s$ but not $X_s$, fitting, or ECE. $J_s$ includes all observed source units, including zero-pass units; $F_s$ counts positive-pass units.

Before approval, select from $0.60/0.70/0.80/0.90$ on the fixed authorized selection pool, maximizing coverage subject to pooled $L > 0.90$ and the count/variation requirements below. Ties favor higher thresholds; no qualifying threshold disables writing. Selection and report coverage are

$$C_{\text{select}}(\tau) = \frac{N_{\text{threshold pass}}(\tau)}{N_{\text{authorized}}}, \qquad C_{\text{report}} = \frac{N_{\text{final admitted}}}{N_{\text{eligible}}}. \tag{10}$$

Both populations retain unusable scores in their denominators. Threshold and maps freeze before quality-partition approval.

**Source-unit bounds and tests.** Each of 10,000 draws samples $J_s$ complete quality units with replacement, preserving candidates and labels. $L_s$ is the 500th ordered $\sum_f X^*_{sf} / \sum_f A^*_{sf}$ (one-sided 5%), without interpolation. Zero-pass units contribute zero to both sums. Zero-denominator cases remain logged, without deletion or replacement; nonfinite $L_s$ fails quality. The test of $H_{0s} : p_s \leq 0.90$ against $H_{1s} : p_s > 0.90$ uses

$$\begin{gathered} U_f = X_{sf} - 0.90 A_{sf}, \qquad T_s = \frac{\sqrt{J_s}\,\bar{U}}{s_U}, \quad U^*_f = \epsilon_f (U_f - \bar{U}), \quad T^*_s = \frac{\sqrt{J_s}\,\overline{U^*}}{s_{U^*}}, \\ b_s = \sum_{b=1}^{B} \mathbf{1}[T^*_{sb} \geq T_s], \quad p^{\text{raw}}_s = \frac{b_s + 1}{B + 1}, \qquad B = 100{,}000. \end{gathered} \tag{11}$$

Means and sample variances use all $J_s$ units; independent Rademacher signs are equiprobable $\pm 1$. $b_s$ counts one-sided tails. Degenerate draws are logged and counted in the tail. Bounds and wild tests are separate approximate procedures, not mutual inversions. Draw counts improve Monte Carlo precision and Holm addresses multiplicity; neither makes the base test exact.

Approval requires a usable fit, $F_s \geq 50$, $A_s \geq 100$, $L_s > 0.90$, and nondegenerate variation. No passes, all-success labels, identical positive-pass unit rates, zero variance, or nonfinite statistics fail quality. Failure sets $p^{\text{gate}}_s = 1$; otherwise $p^{\text{gate}}_s = p^{\text{raw}}_s$. Holm covers all 24 gate values before masking the three `precedes` cells:

$$\text{enabled}_s = \text{authorized}_s \wedge \text{policyOpen}_s \wedge \text{fitUsable}_s \wedge \text{qualityPass}_s \wedge (p^{\text{Holm}}_s \leq 0.05). \tag{12}$$

The source-unit policy leaves 21 policy-open and 12 enabled cells. Approval records retain counts, statistics, decisions, RNG, and seed.

**Counting units under document clustering.** Source-unit $F_s \geq 50$, candidate-count $A_s \geq 100$, and source-unit $L_s > 0.90$ gates remain. Document inference resamples complete parent documents with replacement and uses one wild-test sign per document. $F_s^{\mathrm{doc}}$ counts documents with threshold passes; no $F_s^{\mathrm{doc}} \geq 50$ gate applies. D additionally requires $L_s^{\mathrm{doc}} > 0.90$, recomputed 24-cell Holm correction, and remaining quality checks; U uses source-unit inference. The 19 guideline families split 4/5/5/5, leaving at most five quality documents per cell; a 50-document gate would close all guideline cells. These small-cluster diagnostics are not independent evidence of relation support. U/D enable 12/10 cells: guideline `associated_with`/`treated_with` close under D at 0.896/0.895 (Table D.1). Table 10 reports downstream effects.

**Mapping and source-concentration diagnostics.** Maps record joint-cell → relation → source-class → global fallback, reason, fitting documents and units, sample size, knots, and outputs. Prediction uses right-continuous PAVA steps with endpoint clamping. Source concentration is

$$w_{sf} = A_{sf}/A_s, \qquad \mathrm{HHI}_s = \sum_f w_{sf}^2, \qquad F_{\mathrm{eff},s} = 1/\mathrm{HHI}_s. \tag{13}$$

These quantities summarize source-unit concentration through $\max_f w_{sf}$, HHI, and effective source-unit count.

Table D.1: All 24 quality-check cells and document-cluster sensitivity at the locked threshold. $F_s$ counts source units with threshold-passing candidates; $A_s, X_s$ count threshold passes and reference successes before activation. $L_s$ is the source-unit one-sided support bound and $L_s^{\mathrm{doc}}$ the document-cluster bound. U/D use source-unit/document-cluster inference with source-unit count gates and full 24-cell Holm correction; D adds the document bound, without a $F_s^{\mathrm{doc}} \geq 50$ gate (Appendix D.1).

| Source | Relation | $F_s$ | $A_s$ | $X_s$ | $L_s$ | $L_s^{\mathrm{doc}}$ | U | D |
|---|---|---|---|---|---|---|---|---|
| `dataset_feedback` | `has_finding` | 118 | 1,793 | 1,688 | 0.928 | 0.920 | yes | yes |
| `dataset_feedback` | `associated_with` | 93 | 1,187 | 1,109 | 0.917 | 0.910 | yes | yes |
| `dataset_feedback` | `indicated_for` | 84 | 951 | 903 | 0.934 | 0.928 | yes | yes |
| `dataset_feedback` | `contraindicated_for` | 61 | 405 | 382 | 0.916 | 0.908 | yes | yes |
| `dataset_feedback` | `treated_with` | 108 | 1,512 | 1,421 | 0.924 | 0.916 | yes | yes |
| `dataset_feedback` | `measured_by` | 69 | 674 | 626 | 0.906 | 0.901 | yes | yes |
| `dataset_feedback` | `precedes` | 44 | 291 | 263 | 0.865 | 0.846 | no | no |
| `dataset_feedback` | `causes` | 47 | 149 | 136 | 0.866 | 0.843 | no | no |
| `authorized_note` | `has_finding` | 86 | 1,075 | 998 | 0.905 | 0.901 | yes | yes |
| `authorized_note` | `associated_with` | 64 | 731 | 657 | 0.865 | 0.846 | no | no |
| `authorized_note` | `indicated_for` | 59 | 576 | 528 | 0.889 | 0.866 | no | no |
| `authorized_note` | `contraindicated_for` | 51 | 241 | 221 | 0.878 | 0.850 | no | no |
| `authorized_note` | `treated_with` | 73 | 879 | 812 | 0.897 | 0.876 | no | no |
| `authorized_note` | `measured_by` | 47 | 368 | 330 | 0.862 | 0.838 | no | no |
| `authorized_note` | `precedes` | 29 | 138 | 118 | 0.792 | 0.756 | no | no |
| `authorized_note` | `causes` | 24 | 85 | 68 | 0.721 | 0.674 | no | no |
| `guideline` | `has_finding` | 103 | 1,302 | 1,243 | 0.939 | 0.930 | yes | yes |
| `guideline` | `associated_with` | 79 | 876 | 812 | 0.902 | **0.896** | yes | **no** |
| `guideline` | `indicated_for` | 89 | 1,038 | 982 | 0.930 | 0.921 | yes | yes |
| `guideline` | `contraindicated_for` | 67 | 627 | 593 | 0.925 | 0.916 | yes | yes |
| `guideline` | `treated_with` | 71 | 768 | 711 | 0.902 | **0.895** | yes | **no** |
| `guideline` | `measured_by` | 49 | 411 | 374 | 0.881 | 0.854 | no | no |
| `guideline` | `precedes` | 34 | 172 | 153 | 0.834 | 0.799 | no | no |
| `guideline` | `causes` | 41 | 198 | 178 | 0.851 | 0.818 | no | no |

Failed count/support gates supply one to the full 24-cell Holm correction. Of 24 cells, 10 remain enabled and 12 remain closed; two change from enabled to closed and none opens. The 21-cell policy-open mask is unchanged. Source-unit activation enables six dataset-feedback, one authorized-note, and five guideline cells; document clustering retains six, one, and three, respectively.

The source-unit policy has nine quality-closed cells; document clustering closes two more. Three `precedes` cells are policy-closed. Table 28's authorized-note bound 0.89 uses document-disjoint reporting data; activation uses this quality-check partition.

Table 10: Measured downstream clustering-policy sensitivity on MedQA. Effects and intervals are in percentage points. The last row gives document-cluster minus source-unit policy changes; the CIs refer to Full minus Generic within each policy.

| Admission policy | Full AULC | Generic AULC | Full−Generic | 95% CI |
|---|---|---|---|---|
| Source-unit | 86.80 | 85.92 | $+0.88$ | $[+0.52, +1.24]$ |
| Document-cluster | 86.62 | 85.81 | $+0.81$ | $[+0.42, +1.20]$ |
| Policy change | $-0.18$ | $-0.11$ | $-0.07$ | — |

Table D.2: Isotonic calibration maps for the 12 source-unit-policy enabled cells. Input score breakpoints $x$ and output probabilities $y$ correspond in order.

| Relation | Fit level | Input breakpoints $x$ | Output probabilities $y$ |
|---|---|---|---|
| **dataset_feedback** | | | |
| has_finding | joint | 0.04, 0.16, 0.29, 0.43, 0.57, 0.71, 0.84, 0.93 | 0.07, 0.19, 0.34, 0.51, 0.67, 0.81, 0.90, 0.96 |
| associated_with | joint | 0.06, 0.19, 0.33, 0.48, 0.63, 0.78, 0.90 | 0.09, 0.23, 0.39, 0.56, 0.72, 0.86, 0.94 |
| indicated_for | joint | 0.05, 0.17, 0.31, 0.45, 0.59, 0.73, 0.86, 0.94 | 0.08, 0.21, 0.37, 0.54, 0.70, 0.83, 0.92, 0.97 |
| contraindicated_for | relation | 0.07, 0.21, 0.36, 0.52, 0.68, 0.83, 0.92 | 0.11, 0.25, 0.42, 0.59, 0.75, 0.88, 0.95 |
| treated_with | joint | 0.05, 0.18, 0.32, 0.47, 0.62, 0.76, 0.88, 0.95 | 0.08, 0.22, 0.38, 0.55, 0.71, 0.84, 0.92, 0.97 |
| measured_by | relation | 0.08, 0.23, 0.39, 0.55, 0.71, 0.86, 0.94 | 0.12, 0.27, 0.44, 0.61, 0.77, 0.89, 0.95 |
| **authorized_note** | | | |
| has_finding | source class | 0.09, 0.25, 0.42, 0.58, 0.74, 0.88, 0.95 | 0.13, 0.29, 0.46, 0.63, 0.79, 0.91, 0.96 |
| **guideline** | | | |
| has_finding | joint | 0.03, 0.15, 0.28, 0.42, 0.56, 0.70, 0.83, 0.92 | 0.06, 0.18, 0.33, 0.50, 0.66, 0.80, 0.90, 0.96 |
| associated_with | relation | 0.06, 0.20, 0.35, 0.50, 0.65, 0.79, 0.90 | 0.09, 0.24, 0.40, 0.57, 0.73, 0.86, 0.94 |
| indicated_for | joint | 0.04, 0.16, 0.30, 0.44, 0.58, 0.72, 0.85, 0.93 | 0.07, 0.20, 0.36, 0.53, 0.69, 0.82, 0.91, 0.96 |
| contraindicated_for | joint | 0.05, 0.19, 0.34, 0.50, 0.66, 0.81, 0.91 | 0.08, 0.23, 0.41, 0.58, 0.74, 0.87, 0.94 |
| treated_with | relation | 0.07, 0.22, 0.38, 0.54, 0.70, 0.85, 0.93 | 0.11, 0.26, 0.43, 0.60, 0.76, 0.88, 0.95 |

For $K$ ordered breakpoints, prediction is the right-continuous step $f(x) = y_j$ for $x_j \le x < x_{j+1}$ ($j < K$), with $f(x) = y_1$ below $x_1$ and $f(x) = y_K$ at or above $x_K$.

Table D.3: Source-unit populations and threshold selection. Panel A covers the 12 source-unit-policy enabled cells; panel B shows global-threshold selection.

**A. Source units**

| Source | Relation | $F_s$ | $J_s$ |
|---|---|---|---|
| F | has_finding | 118 | 132 |
| F | associated_with | 93 | 105 |
| F | indicated_for | 84 | 96 |
| F | contraindicated_for | 61 | 72 |
| F | treated_with | 108 | 120 |
| F | measured_by | 69 | 81 |

**A. Source units (continued)**

| Source | Relation | $F_s$ | $J_s$ |
|---|---|---|---|
| N | has_finding | 86 | 99 |
| G | has_finding | 103 | 115 |
| G | associated_with | 79 | 88 |
| G | indicated_for | 89 | 95 |
| G | contraindicated_for | 67 | 73 |
| G | treated_with | 71 | 80 |

**B. Threshold selection**

| Thresh. | $C_{\text{select}}$ | $L_{\text{select}}$ | Pass | Sel. |
|---|---|---|---|---|
| 0.60 | 0.71 | 0.897 | no | no |
| 0.70 | 0.63 | 0.928 | yes | yes |
| 0.80 | 0.48 | 0.941 | yes | no |
| 0.90 | 0.29 | 0.955 | yes | no |

F: dataset feedback; N: authorized note; G: guideline. $J_s$ includes zero-pass source units; $F_s$ counts positive-pass units. Table D.2 gives the maps and fitting levels. Selection uses pooled $L_{\text{select}} > 0.90$; 0.70 maximizes eligible coverage. Quality bounds $L_s$ use the separate quality partition.

# E Evidence Parity and Baseline Fidelity

The adapted baselines in Table 11 share generator, decoding, context, hardware, and visible assertions with polarity, scope, patient ID, source/span, and validity. Graphs index nodes/edges; text verbalizes the same assertions. Graph passages group source-coherent assertions into non-overlapping 256-wordpiece units; Vector RAG follows Appendix H. Raw feedback/option labels are excluded. Recommended and depth/candidate-count/ native-routing variants use common validation, then freeze under native ranks and shared caps; construction and selection follow Appendix A's 50%/100% rule.

**Matched implementations and selection.** Adaptive-RAG keeps no/single/iterative routes, BM25 counts 5/6, and 15 passages; test-derived silver validation is disabled. MedGraphRAG keeps 0.5/0.6 merge thresholds and top-1 graph selection with shared generator/MedCPT. HippoRAG 2 fixes synonym threshold 0.8, 200 candidates, and QA top-$k = 5$. Its selected linking $k = 5$, PPR propagation probability (damping) 0.15, and passage-node weight 0.80 are frozen for evaluation. KEDKG fixes REBEL length 216, three beams/sequences, top-1 candidates, tuning detector threshold.

Table 11: Baseline revisions and validation choices, frozen on 28 August 2026 under the common evidence/generator/decoding/budget contract. Links resolve immutable commits.

| Method | Implementation status | Pinned source | Validation-only degrees of freedom |
|---|---|---|---|
| Adaptive-RAG | official commit; matched corpus/backbone | official revision | T5-large route-classifier epoch in {15,20,25,30,35}; retrieval counts fixed |
| Selective CoT | paper-specified matched reimplementation; no author code | arXiv v1 | none after one prompt is fixed; common stochastic answer decoding and output cap |
| MedGraphRAG | official commit; matched-backbone adaptation | official revision | none; released merge thresholds and top-graph rule fixed |
| HippoRAG 2 | official main commit; no HippoRAG-2 release tag | official revision | linking top-$k$, PPR damping, and passage-node weight |
| KEDKG | official commit; matched stream/backbone | official revision | entity/relation detector threshold only; edit batch fixed by stream protocol |
| Confidence-aware KG | paper-specified matched confidence-filtering control; no author code | arXiv v1 | confidence threshold in {0,.1,...,1}; top-$k$ is not jointly tuned |

Selective-CoT follows Appendix D; confidence-aware KG applies its locked threshold and omits audit/conflict/EMA/PPR.

Pre-enumerate at most 27 settings per method under shared generator/token caps; latency breaks ties. Tuning and final blinded pooled Recall@20/nDCG, linking, and static-support audits use disjoint items/source families. Always-Deep anchors full-compute accuracy/escalation/latency comparisons.

### E.1 Online and Copied-state Component Controls

**Full-stream write-rule interventions.** Each branch evolves private memory over 25 paired 1,000-item MedQA streams, sharing initialization, order, generator/decoding, ACCESS, and READ settings. Full is immediate calibrated (Table 2B). The four gate ablations change one ADMIT rule: calibration, source-relative scope, conflict isolation, or source checking. READ patient compatibility stays active. Table 3A reports complete-history effects with four-test Holm correction; historical reconstruction controls below are state-conditional.

**Budget-matched generic recheck.** To compare targeted scope checking with general rechecking, Generic replaces only ADMIT's separate source-relative scope check. Full and Generic share the complete question, feedback, assertion, and scope, with matched call/token caps and no retries. Appendix D specifies both frozen prompts, candidate selection, and acceptance parsing. Full/Generic/No scope share 25 paired 1,000-item streams, one binary-scored answer per item, and the defined 100-item-window AULC. Final uses the last 200 items (0.5-point steps per trajectory), totaling 4,354/4,323/4,293 correct out of 5,000 per branch. Table 3B uses a matched Full reference and intervals separate from Panel A's four-test family; its Full–Generic Final interval includes zero. Table 12A summarizes the shared caps. The scope diagnostic selects assertions from actual runs, retaining an intact version and modifying or deleting necessary conditions to form each pair. Acceptance measures only the separate scope or Generic call, not complete ADMIT admission (Table 12B–D).

**No ADMIT scope veto.** This branch bypasses only the separate, model-based ADMIT scope decision:

$$\mathrm{keep}_{\text{no-scope}}(e) = \mathbf{1}[r(e) \geq 0.70 \ \wedge \ \mathrm{valid}(e) \ \wedge \ \neg\,\mathrm{conflict}(e)]\,.$$

Extraction still requests all required conditions and stores scope unchanged. The verifier/prompt, calibration/threshold, source/approved-cell, schema, entity, time, and conflict checks remain fixed. READ matches patient ID and all ten scope fields, but can test only retained conditions. Otherwise-eligible assertions can publish despite failing the separate scope-check decision; the contrast estimates this veto's effect over complete histories with other gates active.

**Raw-confidence ablation.** Raw confidence replaces Full's isotonic $\widehat{r}(e)$ with BF16 MedGemma's forced-choice $r_{\mathrm{raw}}(e)$, matching write coverage on validation. Only ADMIT uses the raw score; READ continues to rank dynamic assertions with the same calibrated reliability as Full. Neither uses extractor scores. Verifier/prompt, schema, scope, conflict, and source gates remain fixed.

Table 12: Budget-matched rechecking and paired scope diagnostics. Panels B–D use 100 pairs from run-extracted assertions, each with an intact and a condition-modified version. Acceptance is the check-call decision, not full ADMIT admission. Rate differences and CIs are in percentage points; exact McNemar tests are two-sided and unadjusted.

**A. Additional-check budget**

| Metric | Full | Generic recheck |
|---|---|---|
| Maximum calls per candidate / feedback item | 1 / 8 | 1 / 8 |
| Input/output token caps | 4,096 / 32 | 4,096 / 32 |
| Extra retries | 0 | 0 |

**B. Conditions: erroneous acceptance and intact retention**

| | | Broken accepted ↓ | | Intact retained ↑ | |
|---|---|---|---|---|---|
| Attribute | Pairs | Full | Generic | Full | Generic |
| Age | 10 | 0/10 | 2/10 | 10/10 | 9/10 |
| Sex | 10 | 1/10 | 2/10 | 10/10 | 10/10 |
| Pregnancy status | 10 | 0/10 | 3/10 | 10/10 | 9/10 |
| Weight | 10 | 0/10 | 1/10 | 10/10 | 10/10 |
| Renal function | 10 | 1/10 | 3/10 | 10/10 | 9/10 |
| Disease | 10 | 1/10 | 2/10 | 10/10 | 10/10 |
| Allergy | 10 | 1/10 | 2/10 | 10/10 | 9/10 |
| Care setting | 10 | 0/10 | 2/10 | 10/10 | 10/10 |
| Concomitant medication | 10 | 0/10 | 2/10 | 9/10 | 10/10 |
| Immunization status | 10 | 0/10 | 1/10 | 9/10 | 9/10 |
| **Total** | **100** | **4/100** | **20/100** | **98/100** | **95/100** |

**C. Paired effects**

| Endpoint | Full | Generic | Full−Generic | 95% CI | Raw $p$ |
|---|---|---|---|---|---|
| Condition-error acceptance ↓ | 4.00% | 20.00% | $-16.00$ | $[-24.00, -9.00]$ | 0.000145 |
| Intact-assertion retention ↑ | 98.00% | 95.00% | $+3.00$ | $[-1.00, +8.00]$ | 0.375000 |

**D. Paired acceptance counts**

| Assertion type | Both | Full only | Generic only | Neither | Total |
|---|---|---|---|---|---|
| Condition-broken | 3 | 1 | 17 | 79 | 100 |
| Condition-intact | 94 | 4 | 1 | 1 | 100 |

Table E.4: Coverage-matched operating points for the calibration ablation.

| Variant | Coverage | Threshold |
|---|---|---|
| Full (calibrated) | 0.632 | 0.70 |
| Raw confidence | 0.631 | 0.738 |

The raw-confidence variant uses the raw verifier score only for ADMIT and uses a threshold of 0.738 to match coverage. READ retains calibrated reliability; verifier, prompt, schema, scope, conflict, and source checks stay fixed.

**Three source checks.** **(1) Span validation:** the frozen parser checks contiguous, byte-aligned support, including both endpoints, against retained text: the condition-specific perturbed view for feedback corruption (Appendix M.1), otherwise the retained source version. Completed checks reject mismatches; exceptions or unavailable checks allow candidates to proceed. **(2) Semantic verification:** MedGemma produces $r_{\text{raw}}$ from entailment under retained question conditions. **(3) Source-authorization gate:** ADMIT blocks publication/retrieval unless (i) the family is authorized within `dataset_feedback`, `authorized_note`, or `guideline`; (ii) the source has version metadata and is outside frozen evaluation/future feedback; (iii) source ID, byte-span fields, observation time, and commit record exist; (iv) patient-level IDs match source metadata and population IDs are null; and (v) the source is neither a generated answer treated as truth nor unversioned web text. This field-existence gate does not establish version completeness or successful span reconstruction. Online committed/released feedback is eligible. No source-auth gate removes layer 3, including source-dependent approved-cell admission. Full retains its 24-cell rules. Span/semantic checks, the frozen score mapping, relation-policy restrictions, schema/link, scope, time, conflict checks, and prediction-before-feedback visibility remain active.

**ADMIT-boundary injection test.** Question: antibiotic for adult cellulitis with penicillin allergy; released answer: "Clindamycin." Injected assertion: `(cellulitis, treated_with, clindamycin)`, population-level with adult age scope, penicillin-allergy scope, and null patient

ID. The coded field is `allergy_set={SNOMED CT:91936005}`, denoting *Allergy to penicillin (finding)* (SNOMED International, n.d.). The patient's resolved allergy set contains this finding, so the allergy predicate passes. An empty or unresolved patient allergy set fails; other unconstrained fields are explicitly `unrestricted`. Source: `model_generated_explanation`, `gen-exp-2026-08-15`. Normal extraction excludes this source; the test injects it at ADMIT. Span/entailment pass: $r_{\text{raw}} = 0.91$. Its unregistered class uses the frozen global isotonic map, $\widehat{r} = 0.88 \geq 0.70$, rather than stratified maps. Schema/scope pass; no conflict is detected. Full rejects, preventing publication/retrieval. No source-auth gate bypasses source authorization and source-cell admission, permitting publication and later retrieval.

**Copied-state controls.** Tables 13/20 use 1,273 frozen MedQA items, five stochastic answers/item, $K = 20$, $K_d = 5$, and 1,024 evidence tokens. Write variants reconstruct disposable memory from shared construction/historical proposals before QA; current evaluation feedback is excluded. READ variants share Full memory. Routing forces paths or $w_H \in \{0, 1\}$ and reselects only the validation threshold; Deep keeps its retriever/prompt. Intervention and reference metrics follow Appendices H and L. The separate corruption study uses 500 Deep items and greedy answers (Table 29).

Table 13: MedQA component controls with shared historical proposal inputs. Accuracy and pooled-reference recall use fixed budgets; paired accuracy effects appear in Table 20.

| Configuration | Acc. ↑ | Pooled R@20 ↑ | Read ECE ↓ | p50 s ↓ |
|---|---|---|---|---|
| Static graph only | 76.6 | 0.612 | 0.057 | 1.02 |
| Dynamic graph only | 79.9 | 0.744 | 0.041 | 1.08 |
| No semantic rerank; links retained | 79 | 0.701 | 0.049 | 1.11 |
| No confidence rerank | 79.5 | 0.721 | 0.043 | 1.19 |
| Dense retrieval; no links | 77.6 | 0.612 | 0.062 | 0.74 |
| Raw verifier confidence | 80.1 | 0.734 | 0.039 | 1.23 |
| No source-auth gate | 79.2 | 0.719 | 0.063 | 1.20 |
| No conflict isolation | 79.6 | 0.727 | 0.039 | 1.18 |
| **EvoMedAgent (dual, full)** | 82.1 | 0.781 | 0.028 | 1.27 |

**Write ECE and its evaluation population.** Resolved joint-reference labels and bins follow Appendix L. $N$ counts ADMIT evaluation candidates; static/dynamic rows use source subsets.

| Configuration | ECE | $N$ | Configuration | ECE | $N$ |
|---|---|---|---|---|---|
| Full | 0.031 | 8921 | No confidence rerank | 0.031 | 8921 |
| Static graph only | 0.044 | 5913 | Raw verifier confidence | 0.039 | 8921 |
| Dynamic graph only | 0.036 | 3008 | No source-auth gate | 0.031 | 8921 |
| No semantic rerank; links retained | 0.031 | 8921 | No conflict isolation | 0.036 | 9117 |

READ-only changes preserve candidates, scores, and labels. Raw confidence uses pre-isotonic MedGemma probabilities on Full's resolved cohort, retaining labels, weights, and bin edges. No source-auth gate also keeps Full's calibrated scores and resolved cohort. No conflict adds 196 otherwise eligible contested candidates with resolved labels, keeping common scores/labels and the 15 bin edges fixed while recomputing bin counts/means. Joint/source-support targets and the separate conflict-detection task follow Appendix L and Table 6. Pooled ECE permits signed-error cancellation between source subsets.

### E.2 Knowledge-source Inventory

Initial sources, eligible by 28 August 2026, exclude calibration and future feedback. A split-blind steward audits released partitions and tombstones answer-bearing units for all methods; the fixed mask yields clean-subset contrasts (Table 6). Static evidence is PrimeKG v2.1 (https://doi.org/10.7910/DVN/IXA7BM). Licensed UMLS 2026AA aliases are English, unsuppressed, CUI-deduplicated, and matched without fuzzy expansion. Initial payloads exclude benchmark questions/options/keys and generated explanations.

**Source reconstruction.** Membership and exclusions freeze before evaluation, using versions through 28 August 2026. Guideline-A draws only from Table 16's six documents/sections, initially permitting `indicated_for`/`contraindicated_for`. Extraction retains recommendation

Table 14: Initial frozen/online evidence: static edges, construction questions, and dynamic assertions. EHR transfer starts with empty dynamic memory.

| Source ID | Version | Role/count unit | Count |
|---|---|---|---|
| PrimeKG | v2.1 | Static graph edges; authorized | 120,000 |
| UMLS | 2026AA | Ontology for entity linking; authorized | – |
| Dataset-feedback | – | MedMCQA construction questions; authorized | 16,024 |

| Dynamic source | Release | Before exclusion | Excluded | Retained |
|---|---|---|---|---|
| CDC/WHO guideline collection | G01–G06 | 4,320 | 120 | 4,200 |
| MIMIC-IV-Note discharge summaries | v2.2 | 2,875 | 75 | 2,800 |

Both dynamic sources are authorized. Guideline-A denotes G01–G06 in Table 16; Note-B denotes MIMIC-IV-Note v2.2 discharge summaries from construction patients only. All three numerical columns in the dynamic-source panel count assertions. Extraction scope and final source–relation admission are recorded separately in Appendix E.2. Construction feedback supports the eight schema relations and is counted in questions; its assertion count is separate. "–" denotes an unenumerated quantity.

Table 15: MedQA decontamination: EvoMedAgent minus Always-Deep, applying the fixed mask to unchanged frozen predictions. Paired intervals/raw $p$ are unadjusted.

| Analysis | Unique items | $\Delta$ Acc. (pp) | 95% CI | Raw $p$ |
|---|---|---|---|---|
| Full split | 1273 | +1.6 | $[+0.3, +2.7]$ | 0.011 |
| Clean subset | 1,146 | +1.4 | $[+0.1, +2.7]$ | 0.0150 |

strength, population, polarity, and time; it rejects assertions whose schema conflates cautions, weak recommendations, or explicit contraindications.

Table 16: Guideline-A document roster and eligible extraction sections.

| ID | Official document | Release | Eligible sections |
|---|---|---|---|
| G01 | CDC: Sexually Transmitted Infections Treatment Guidelines | 2021 | Syphilis, chlamydia, gonorrhea, pelvic inflammatory disease, and their special-population guidance |
| G02 | CDC: U.S. Medical Eligibility Criteria for Contraceptive Use | 2024 | Appendices B–D, including category definitions and clarifications |
| G03 | CDC/NTCA: Guidelines for the Treatment of Latent Tuberculosis Infection | 2020 | Recommended regimens and Other Considerations |
| G04 | WHO: Guideline for the pharmacological treatment of hypertension in adults | 2021 | Treatment initiation, drug classes, combination therapy, targets, and follow-up |
| G05 | WHO: Guidelines on meningitis diagnosis, treatment and care | 2025 | Diagnosis, antimicrobial treatment, and adjunctive treatment |
| G06 | WHO: Recommendations on maternal and newborn care for a positive postnatal experience | 2022 | Maternal and newborn care recommendations in Chapter 3 |

**Authorized-note corpus.** Note-B uses `discharge.csv.gz` from MIMIC-IV-Note v2.2. Only the 462 construction patients among 962 EHRNoteQA patients contribute; all 500 evaluation patients' notes are excluded, including other admissions. EHRNoteQA `patient_id` joins MIMIC `subject_id`; unique notes sort by `charttime`, `hadm_id`, then `note_id`. The first three available summaries freeze before extraction. Initial `has_finding`/`treated_with` assertions retain clinical conditions and source `subject_id` as `patient_id`.

**Calibration sources and isolation units.** Table E.3's catalogue freezes before reference scoring and excludes initial G01–G06 families. Revisions, translations, mirrors, derivatives, and near-duplicates share a document family and partition. Chapters and recommendation entries stay within that partition but form fitting and inference units; $J_s$ and $F_s$ count units, not documents.

Note calibration excludes all 962 EHRNoteQA patients. PCG64 seed 61003 samples up to 2,000 sorted eligible `subject_id`s without replacement, with up to three notes each as above. Patient and connected near-duplicate groups remain intact. Feedback-calibration families are reserved before initial-feedback selection from the construction budget and excluded from initial memory and evaluation. Seed 61001 partitions intact families into fit/selection/quality-check/report with pooled target proportions 40/20/20/20 across all source classes; individual classes need not match these proportions. The 19 guideline families split 4/5/5/5. Admission follows Appendix D.1's counts, nondegeneracy, lower bound, 24-cell Holm, and fallback quality requirements; insufficient cells stay closed.

Table E.3: Calibration guideline catalogue and document-level partition assignment. All entries have source class `guideline`.

| ID | Document (year) | ID | Document (year) |
|---|---|---|---|
| **Fit** | | **Quality** | |
| C01 | IDSA/ASM Microbiology Lab. (2018) | C11 | KDIGO CKD (2024) |
| C02 | IDSA Skin/Soft Tissue Infection (2014) | C12 | ADA Diabetes Standards of Care (2025) |
| C03 | ATS/IDSA HAP/VAP: VAP sections (2016) | C13 | GOLD COPD (2025) |
| C04 | ATS/IDSA HAP/VAP: HAP sections (2016) | C14 | AHA/ASA Stroke (2021) |
| C05 | AHA Endocarditis (IDSA endorsed) (2015) | C15 | ACG GI Bleeding (2021) |
| **Selection** | | **Report** | |
| C06 | IDSA Prosthetic Joint Infection (2013) | C16 | WHO Malaria (2022) |
| C07 | IDSA Complicated UTI (2025) | C17 | WHO Tuberculosis (2022) |
| C08 | IDSA Candidiasis (2016) | C18 | CDC Rabies (2021) |
| C09 | ATS/IDSA Community-Acquired Pneumonia (2019) | C19 | NICE Sepsis (2024) |
| C10 | ESC Heart Failure (2021) | C20 | IDSA/SHEA Antibiotic Stewardship (2016) |

C03 and C04 denote the VAP and HAP extraction sections of the same 2016 ATS/IDSA guideline. They share one document family and both belong to fit.

Qwen3.5-9B extraction and MedGemma-27B verification preserve schema, prompts, linking, scope, and conflict rules. Initial and subsequent candidates need approved admission plus extraction eligibility; exact duplicates merge provenance. Matched graph/text methods share initial facts, retrieval budgets, and decoding. All 25 EHR transfer histories start from shared PrimeKG and empty dynamic memory, without Note-B, guideline, or construction-feedback preloads; upstream writes follow post-prediction reveals.

### E.3 Constructing and Scoring the Online Streams

Each method answers once per item from private committed memory; scoring precedes scheduled question/answer release, which no-update discards. Weights, prompts, thresholds, and selected settings remain fixed.

**Item order and feedback.** The 1,000 MedQA and 3,500 MedMCQA items exclude initial-memory construction. Order seeds 41001–41025 pair by index with generation-run seeds 51001–51025. Primary PCG64 orders are uniform permutations; shifted orders follow Appendix M.1. Methods share underlying initial memory and order/run IDs; CM masks its visibility. Request seeds include method identity (Appendix K). Frozen and no-update requests share initial memory, Access/Read, generator, prompts, caps, and sampling. Matching complete item–seed inputs fixes options, evidence, and rendered prompts; frozen sampling uses seeds 2027–2031.

Primary streams reveal every correct answer after prediction; nine sensitivity conditions vary availability, correctness, or delay (Figure 5; Appendix M.1). True answers score predictions; only admitted feedback/visibility events change memory. Delayed feedback waits 100 further predictions. No-update retains full initial memory and discards feedback. Post-stream feedback cannot affect scores.

**Controlled masking (CM).** Fix $B = 100$ and initial dynamic transaction/edge counts $(T_{\text{init}}, E_{\text{init}})$. Mask fraction $f(t) = \max(0, 1 - t/B)$ at position $t$. Over positions 1–100, monotonically restore counts toward $(tT_{\text{init}}/B, tE_{\text{init}}/B)$ within indivisible source/edge units. Mask lowest-priority units by source class then canonical edge key, identically across trajectories; route, answer, and score against this memory. CM immediate/delayed queue feedback released during burn-in that passes ADMIT. Before prediction 101, restore full initial memory and apply this queue once in original order. Later writes follow their respective feedback schedules after answer commitment. Historical permutation restores initial memory identically but follows the block schedule below. Table 32 reports observed terminal counts. CM jointly varies evidence availability, content, and timing.

**Historical permutation.** Use 100-position blocks without a burn-in feedback queue. Prediction 101 uses full initial memory before first-block feedback. After each answer at positions 101–200, process one first-block question–answer pair in a fixed shuffled order per trajectory; admissions

become visible at positions 102–201. Later blocks process the preceding block's pairs, each once. No unrevealed future answer enters prediction.

**AULC and Final.** Let $b_t$ be percentage accuracy over the last $\min(t, 100)$ predictions. Then $\mathrm{AULC} = T^{-1} \sum_{t=1}^{T} b_t$. Final uses the last 20% of predictions: 200 for MedQA, 700 for MedMCQA. Whole-stream accuracy weights answers equally; AULC weights their contributions to overlapping windows, including shorter initial windows.

**Paired statistical inference.** The 25 paired-history effects are conditional on the fixed pool/configuration; inference and multiplicity follow Appendix A.1.

**Frozen and fixed-memory comparison.** Frozen five-answer scoring and 25 online trajectories share a 1,000-item pool. Update effects pair online methods (Table 21); fixed-memory checks match IDs, complete seeds, rendered inputs, and $K_0$. Whole-stream accuracy averages 25,000 answers, Final pools the last 5,000, and AULC retains rolling-window weights.

Frozen/common-pool accuracy is 82.1% versus no-update 81.8%; the 0.3-point gap reflects differing generation counts/seeds (Appendix E.3).

Frozen scoring uses generation seeds 2027–2031; online histories use order seeds 41001–41025 and generation seeds 51001–51025. The frozen benchmark requests five answers for each of 1,273 items, including the common 1,000-item subset; each online method requests one answer per item in each of 25 histories.

## F Leakage Audit Details

Before each run, static sources and initial memory undergo a fixed leakage audit. Online duplicate handling follows the separate assertion rules in Appendix B. Unicode compatibility, case, whitespace, and punctuation are normalized while clinical numbers/negation are preserved. Exact matches are removed; Approximate token 5-gram matching nominates lexical duplicates requiring exact Jaccard $\geq 0.80$. Frozen `all-mpnet-base-v2` nominates paraphrases at cosine $\geq 0.90$. Selene Mini's fixed offline rule determines whether candidates independently reveal an answer. Template signatures replace linked concepts with semantic types and numbers with unit-aware slots, treating options as multisets; template hits are excluded only if also judged answer-bearing. Positive hits tombstone the complete source unit for all methods. Thresholds, pair IDs, and decisions are locked before inference.

The frozen model labels pairs duplicate, related but independently answerable, or non-duplicate. Blinded pair decisions define the fixed item-level mask (coverage: Table 6). Full-split and mask-filtered estimates use their respective item populations.

## G Router Analysis

Each item/patient's deterministic Fast/Deep decision is shared by all five answers; Deep rate counts unique items. MedQA workload traces vary the threshold from 0.2 to 0.8 in increments of 0.1, with Always-Fast and Always-Deep as endpoints. Separate ROC and precision–recall analyses use paired labels identifying cases where Deep corrects a Fast error.

Table 17: Fixed-router accuracy and answer-path cost. $n_{\mathrm{score}}$ counts items/patients, each with five separately scored answers; accuracy averages their means, while latency summarizes requests.

| Dataset | $n_{\mathrm{score}}$ | Accuracy | p50 s | p95 s |
|---|---|---|---|---|
| MedQA | 1273 | 82.1 | 1.28 | 2.03 |
| MedMCQA | 4180 | 77 | 1.19 | 1.38 |
| PubMedQA | 500 | 86.6 | 1.72 | 1.72 |
| MMLU-Gene | 100 | 92 | 1.24 | 1.67 |
| MMLU-Clinic | 264 | 91.7 | 1.07 | 1.45 |
| EHRNoteQA | 500 | 78.4 | 1.01 | 1.67 |

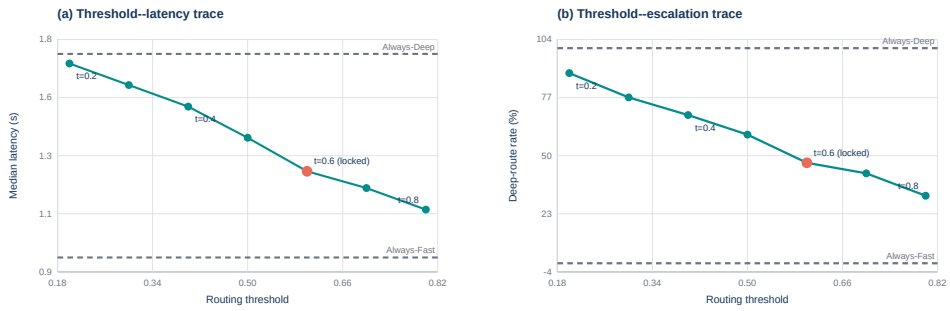


Figure 9: MedQA threshold sweep: median latency and Deep rate. Dashed: Always-Fast/Deep; coral: fixed threshold. Accuracy: Table 17.

**Normalizing the probe and selecting the route.** ACCESS uses two deterministic probes: uncertainty $H(q)$ and linked-concept density $D(q)$. For multiple-choice questions, $H(q)$ is option-probability entropy divided by $\log|\text{options}|$. For open-ended questions,

$$H(q) = \frac{1}{m_q}\sum_{t=1}^{m_q} H(p_t). \tag{14}$$

Here $p_t$ is the full-vocabulary FP32 distribution at temperature one, before filtering. Entropy uses natural logarithms and is computed before greedy selection. The average includes the EOS-producing step: immediate EOS gives $m_q = 1$, and the length cap gives $m_q = 32$. NaN/Inf invalidates the probe and sends the question to Deep.

For density, NFKC normalization preserves numbers, units, comparison operators, negation, and time, without stemming or fuzzy abbreviation expansion. Length is the original, untruncated question's Qwen3.5-9B token count, excluding special tokens for every backbone. Initial buckets follow construction-data length quartiles. Scan from shortest to longest, merging buckets with fewer than 500 construction examples rightward, or leftward if no right neighbor remains, until each reaches 500 or only one remains. Merged buckets share their respective construction-data percentile maps. Count each mention without entity-ID deduplication:

$$D_{\text{raw}}(q) = \sum_{i=1}^{K} w_i, \qquad w_i = \begin{cases} 1, & \text{linked,} \\ 0.5, & \text{unresolved,} \\ 0, & \text{nonconceptual.} \end{cases} \tag{15}$$

For the $N$ construction values $D_i$ in the question's length bucket,

$$\widetilde{D}(q) = \begin{cases} 0, & D_{\text{raw}}(q) = 0, \\ \dfrac{\#\{D_i < D_{\text{raw}}(q)\} + \frac{1}{2}\#\{D_i = D_{\text{raw}}(q)\}}{N}, & D_{\text{raw}}(q) > 0. \end{cases} \tag{16}$$

Normalize $H(q)$ by the ordinary within-bucket midrank to obtain $\widetilde{H}(q)$. An empty mention set has density zero. The ordinary uncertainty midrank does not use the density-specific zero rule. For multiple-choice tasks, minimize validation option negative log likelihood by bounded Brent search in log-temperature space: temperature 0.05–10, at most 200 iterations, tolerance $10^{-7}$. EHR uses temperature one.

Search $w_H \in \{0, 0.25, 0.5, 0.7, 0.9, 1\}$ and $\tau \in \{0.2, 0.3, \ldots, 0.8\}$. Among configurations passing Appendix A.1's paired-unit $L_c > -0.005$ screen, minimize mean latency; break ties by lower Deep rate, higher threshold, then lower $w_H$. MedQA locks $\tau = 0.60$. The screen applies to overall mean accuracy, not Fast/subgroup/item guarantees. Selection failure triggers Section 3.2's logged Always-Deep fallback. Temperature, buckets, maps, and configuration freeze before evaluation.

### G.1 FINAL OPERATING CONFIGURATION

Table G.2 gives the final Qwen3.5-9B probe settings; generation temperature is specified separately. ADMIT's threshold is 0.70. Native HippoRAG 2 uses linking $k = 5$, propagation/damping 0.15 (restart 0.85), and passage weight 0.80; controlled retrieval uses restart 0.15 (Appendix H). Configuration records include merged length ranges/maps, selection-set ID, mean latency, fallback, and versions of code, configuration, model, tokenizer, and index.

Table G.2: Qwen3.5-9B frozen-evaluation configuration: $w_H = 0.50$, $\tau = 0.60$, and Deep when $C(q) > \tau$. EHR transfer/masking instead use Always-Deep. Counts/Deep rates are frozen-evaluation values; $L_c$ is the selection bound on the 0–1 scale; p50 is in seconds.

| Dataset | Items | Deep items | Deep (%) | $T_{\text{probe}}$ | $L_c$ | p50 |
|---|---|---|---|---|---|---|
| MedQA | 1273 | 594 | 46.7 | 0.85 | -0.001 | 1.28 |
| MedMCQA | 4180 | 1848 | 44.2 | 0.85 | -0.002 | 1.19 |
| PubMedQA | 500 | 207 | 41.4 | 0.85 | -0.003 | 1.72 |
| MMLU-Gene | 100 | 41 | 41.0 | 0.85 | -0.004 | 1.24 |
| MMLU-Clinic | 264 | 114 | 43.2 | 0.85 | -0.002 | 1.07 |
| EHRNoteQA | 500 | 207 | 41.4 | 1.00 | -0.003 | 1.01 |

Table G.3: Master seeds by purpose. Online generation-run seeds pair by trajectory index; request-level streams remain method-specific.

| Purpose | Seed | Role |
|---|---|---|
| Frozen generation | 2027–2031 | Five answers per item |
| Online order | 41001–41025 | 25 trajectories |
| Online generation | 51001–51025 | Paired with order by index |
| Family split | 61001 | Pooled target 40/20/20/20 |
| ADMIT wild | 61002 | 100,000 draws |
| Note calibration | 61003 | 2,000 subject IDs |
| EHR pilot | 314159 | 100 of 462 patients |

Seed records contain `run_id`, `trajectory_id`, `purpose`, `rng_algorithm`, `seed`, `input_order_id`, and `output_manifest_id`; purposes cover split, order, visibility, corruption, generation, bootstrap, and test randomization. Items retain visibility, corruption, and substituted option; EHR records add upstream/downstream orders, generation stream, and five generation-index seeds. Shared masks/orders use shared IDs. Assignments link `query_id`, `domain_id`, `original_category`, `mapping_rule_version`, and `ambiguity_resolution`; $D_1$–$D_4$ rules specify names, category sources, multi-label priority, unassignable handling, membership, and counts. Configurations and assignments freeze before evaluation.

Table G.4: Final MedQA routing configurations for the other backbones. Deep counts use all 1,273 evaluation items. Mode HD combines entropy and density; mode D uses density alone. Hosted GPT backbones do not run an entropy probe. No row uses the Always-Deep fallback.

| Backbone | Mode | $w_H$ | $\tau$ | $T_{\text{probe}}$ | Deep items | Deep (%) | Fallback |
|---|---|---|---|---|---|---|---|
| Qwen3.5-0.8B | HD | 0.50 | 0.60 | 0.85 | 562 | 44.1 | no |
| MedGemma 27B | HD | 0.50 | 0.60 | 0.85 | 577 | 45.3 | no |
| Ministral 3 8B | HD | 0.50 | 0.60 | 0.85 | 558 | 43.8 | no |
| gpt-oss-20B | HD | 0.50 | 0.60 | 0.85 | 566 | 44.5 | no |
| Qwen3.8-27B-FP8 | HD | 0.50 | 0.60 | 0.85 | 574 | 45.1 | no |
| GPT-5.6 Luna | D | 0.00 | 0.60 | off | 544 | 42.7 | no |
| GPT-5.6 Terra | D | 0.00 | 0.60 | off | 544 | 42.7 | no |
| GPT-5.6 Sol | D | 0.00 | 0.60 | off | 544 | 42.7 | no |

**Executable routing interventions.** Always-Fast/Deep force routes; entropy-only sets $w_H = 1$ and density-only $w_H = 0$. Each reselects only its threshold on the same non-evaluation split under the full router's accuracy-loss constraint/tie rules. Selective CoT uses its native DIRECT/COT choice; Adaptive-RAG uses its native no/single/iterative retrieval paths. These baselines share the generator, decoding limits, evidence access, and timing procedure. Their escalation rate counts COT choices

or queries with at least one retrieval call, respectively. For EvoMedAgent and its routing ablations, escalation means selecting Deep. The diagnostic oracle chooses Deep for Deep-correct/Fast-incorrect items and Fast otherwise.

**Option-permutation audit.** Complete item/permutation option-probability vectors are restored to canonical indices before argmax-agreement, Jensen–Shannon-divergence, route-flip, and accuracy calculation. Agreement uses these paired vectors; operating points summarize each configuration separately.

Table 18: MedQA paired effects (EvoMedAgent minus comparator). A: six-test Holm family, distinct from frozen comparisons; bold $p \leq 0.05$ and raw $< 0.0001$ denotes a bound. B: paired p50 intervals and p95 differences (Table 4), outside that family.

**A. Accuracy inference**

| Comparator | Δ Acc. (pp) | 95% CI | Raw $p$ | Holm $p$ |
|---|---|---|---|---|
| Always-Fast | +12.1 | [+10.5, +13.7] | $< 0.0001$ | $< 0.0001$ |
| Always-Deep | +1.6 | [+0.3, +2.7] | 0.011 | **0.011** |
| Selective CoT | +11.2 | [+9.8, +12.8] | $< 0.0001$ | $< 0.0001$ |
| Adaptive-RAG | +8.3 | [+7.2, +9.8] | $< 0.0001$ | $< 0.0001$ |
| Entropy only | +7.4 | [+6.1, +8.7] | $< 0.0001$ | $< 0.0001$ |
| Density only | +7.1 | [+5.8, +8.4] | $< 0.0001$ | $< 0.0001$ |

**B. Compute contrasts; brackets give paired 95% CIs**

| Comparator | Δ p50 s | Δ p95 s | Δ escalation rate (pp) |
|---|---|---|---|
| Always-Fast | +0.36 [+0.31, +0.41] | +0.72 | +46.7 |
| Always-Deep | -0.49 [-0.54, -0.44] | -0.83 | -53.3 |
| Selective CoT | -0.03 [-0.07, +0.01] | -0.01 | +3.6 |
| Adaptive-RAG | -0.11 [-0.15, -0.07] | -0.15 | -2.8 |
| Entropy only | -0.06 [-0.10, -0.02] | -0.13 | -4.3 |
| Density only | -0.01 [-0.05, +0.03] | -0.04 | -1.6 |

# H Retrieval-component Controls

Table 4B fixes memory, generator, evidence caps, and relevance audit; audit labels never tune controls. Dense retrieval cosine-ranks all serialized triples. Linking restricts candidates to accepted entities; PPR-only propagates from those seeds without cosine reranking; the combined control links, propagates, and reranks. HippoRAG 2 uses its pinned official passage integration with shared evidence, generator, and context cap.

Table 4B uses $K = 20$, $K_d = 5$, and 1,024 graph tokens, with five scored answers per item (packing: Appendix H.1).

## H.1 Matching Evidence to the Current Patient

READ accepts non-null `patient_id` only when the resolved current `subject_id` matches exactly; null-ID assertions skip identity alone. Patient attributes use the current question and generator-visible pre-feedback record; restrictions come from source-supported scope. Absent, ambiguous, negated, and positive observations remain distinct. Only explicitly `unrestricted` fields bypass checks; unresolved constrained attributes fail.

**Physiological state.** Patient age $P = [p_{\min}, p_{\max}]$ must lie within evidence age $E = [e_{\min}, e_{\max}]$:

$$P \subseteq E \quad \Longleftrightarrow \quad e_{\min} \leq p_{\min} \ \wedge \ p_{\max} \leq e_{\max}.$$

Exact age $p$ uses $P = [p, p]$; $[17, 19]$ fails an age-18-or-older restriction. Weight uses identical containment after kg conversion; open endpoints remain strict. Renal checks require same-type, same-unit eGFR interval containment or membership in explicitly allowed documented CKD stages. Neither substitutes for the other; both must pass when constrained. Sex, pregnancy, and lactation require exact category matches. Pregnancy and lactation may coexist but are never inferred from each other.

**Clinical state and care context.** Active conditions must cover all required SNOMED CT/ICD-10 codes through exact normalization or explicit versioned equivalence, not semantic similarity. Negated, historical, or unresolved conditions cannot satisfy active-condition requirements. Care setting matches exact categories (e.g., inpatient, outpatient, ICU, ED) or an explicit candidate wildcard.

For resolved positive coded patient allergies $A_q$ and candidate-required allergies $A_e$, direct matching requires $A_e \subseteq A_q$. Cross-sensitivity needs an authorized, versioned, direct mapping from a documented patient allergen to the requirement, satisfying its reaction and clinical conditions. Finding-to-allergen conversion is explicit; substance mentions alone do not establish allergy. Symmetric completion, transitive closure, drug-class expansion, and embedding similarity cannot infer cross-sensitivity; missing mappings fail. Empty, negated, or unresolved allergy sets cannot satisfy nonempty requirements; only `unrestricted` evidence skips this check.

**Medication and intervention state.** Current medications use resolved RxNorm ingredient codes. Required co-medications match exact normalized codes or versioned equivalents; joint-use assertions require all partners. Interaction assertions require a match to their alternative partner set or a source-supported interaction between candidate drug and current medication, under stated conditions. The source must specify these alternatives; arbitrary medication overlap is insufficient. Unknown status establishes neither an interaction nor its absence. Immunization matches every required vaccine-code/status pair at current record time; missing or unknown status fails.

All ten fields' declared predicates must pass before ranking, scanning, and budgeting. ADMIT checks condition retention; READ matches stored conditions to the patient. Omitted conditions can escape both checks; neither ensures clinical applicability or treatment safety. `certainty` passes unchanged to the prompt, retaining confirmed/probable/uncertain distinctions.

**Selecting evidence within the budget.** Filter at most 100 nearest-neighbor candidates for visibility, admission, and patient compatibility (Appendix H.1); dynamic edges also require MedCPT cosine $\geq 0.70$. Clip cosine to $[0, 1]$ and score as $0.80\,\mathrm{cosine} + 0.20\,\mathrm{reliability}$, minus 0.05 if neither endpoint is linked to the query (all candidates when no entity links). Sort by score, reliability, then canonical subject, relation, object, polarity, scope, validity, patient ID, certainty, and layer. Scan for 20 edges, skipping further dynamic edges after five. Serialize with the generator's tokenizer; drop whole bottom-ranked edges until within 1,024 tokens, without refilling. Encoder/linking rules follow Appendices A and B.

**How the PPR control works.** PPR propagates normalized accepted entity-link scores through the eligible, patient-compatible graph; no accepted entity yields empty evidence. Forward and typed inverse links use read reliability, with parallel weights summed and outgoing weights normalized. Restart probability is 0.15; dangling nodes use the starting distribution. After at least ten iterations, stop at total absolute change $\leq 10^{-6}$ or iteration 200, marking capped results nonconverged. PPR-only ranks mean endpoint probability, breaking ties by canonical ID. Query–triple plus PPR combines independently zero–one-scaled cosine/PPR scores with weights 0.80/0.20 under shared caps. Passage nodes, answers, and test labels never seed propagation.

Table 20 fixes proposals, queries, encoders, and budgets (Appendix E.1). Static/dynamic-only leaves omitted slots empty; no query–triple match zeros cosine after linking; no confidence rerank fixes read reliability, preserving ADMIT. Dense/no-link reuses Table 4B. Write interventions rebuild disposable histories before QA; Raw confidence and No source-auth gate follow Appendix E.1, with ECE populations defined there. No-conflict adds eligible contested candidates without changing scores/provenance.

**Vector RAG configuration and relevance evaluation.** Frozen Vector RAG uses one record per evidence unit: common static-edge verbalizations and non-evaluation dynamic reference spans. Records over 256 tokenizer tokens become 256-token windows with 32-token overlap; shorter records stay separate. Frozen `all-mpnet-base-v2` exact cosine retrieves top 20 windows, breaking ties by record ID before the common token cap. It uses no entity links/graph edges and shares generator, prompt, decoding, and visibility. Online editors instead follow Appendix I.

Before viewing outputs, the final audit samples 510 test questions by dataset, route, and linked-entity count. Each retriever contributes up to 100 eligible canonical edges per question under native pre-cap

scores; passages expand in passage/canonical-edge order before union deduplication. This pool supplies the common denominator. Method-masked Selene Mini relevance grading (Appendix D) hides verifier, method, route, admission, and rank metadata: 0 irrelevant, 1 supporting context, 2 directly answer-bearing. The audit contains 4621 model-rated pairs. Appendix L defines query-averaged Pooled R@20, its shared denominator, and empty-pool convention.

Table H.2: Per-query retrieval on MedQA. The paired MedQA subset contains 312 of the 510 cross-dataset audit queries, including 18 queries with empty relevant pools.

| | | Relevant hits | | $R$@20 | |
|---|---|---|---|---|---|
| Query | $\lvert G_q \rvert$ | EvoMedAgent | HippoRAG 2 | EvoMedAgent | HippoRAG 2 |
| q001 | 12 | 9 | 8 | 0.750 | 0.667 |
| q002 | 8 | 6 | 5 | 0.750 | 0.625 |
| q003 | 15 | 11 | 10 | 0.733 | 0.667 |
| q004 | 10 | 7 | 6 | 0.700 | 0.600 |
| q005 | 6 | 4 | 4 | 0.667 | 0.667 |

Scoring uses `ranked_top20`, before token packing, with a shared relevant pool for each paired query. Empty pools score zero and remain in the 312-query mean. The complete 510-query audit contains 4,621 relevance judgments.

Table 19: Sealed MedQA paired retrieval effects: EvoMedAgent minus comparator. Accuracy and pooled-reference Recall@20 each form a separate six-test Holm family; bold adjusted $p \leq 0.05$. The MedQA HippoRAG 2 comparison uses the same frozen scores, interval, and raw $p$ as Table 30.

| | Accuracy | | | | Pooled R@20 | | | |
|---|---|---|---|---|---|---|---|---|
| Comparator | $\Delta$ pp | 95% CI | Raw $p$ | Holm $p$ | $\Delta$ pp | 95% CI | Raw $p$ | Holm $p$ |
| Dense retrieval | +4.5 | [+3.2, +5.8] | $< 0.0001$ | $< 0.0001$ | +16.9 | [+14, +19.8] | $< 0.0001$ | $< 0.0001$ |
| Query–triple only | +3.7 | [+2.4, +5] | $< 0.0001$ | $< 0.0001$ | +10 | [+7.6, +12.4] | $< 0.0001$ | $< 0.0001$ |
| PPR only | +3.2 | [+1.9, +4.5] | $< 0.0001$ | $< 0.0001$ | +7.7 | [+5.4, +10] | $< 0.0001$ | $< 0.0001$ |
| Query–triple + PPR | +2.5 | [+1.2, +3.8] | 0.0002 | **0.0006** | +5.3 | [+3.2, +7.4] | 0.0002 | **0.0006** |
| HippoRAG 2 | +2.0 | [+0.8, +3.4] | 0.0021 | **0.0042** | +4 | [+2, +6] | 0.0025 | **0.0050** |
| Confidence-aware KG | +1.8 | [+0.5, +3.1] | 0.0040 | **0.0042** | +2.9 | [+0.9, +4.9] | 0.0040 | **0.0050** |

Table 20: Copied-history effects (Full minus variant); eight-contrast Holm family, bold $p \leq 0.05$. Dense reuses Table 19's scores/CI/raw $p$; its Holm bound is $8p$.

| Variant | $\Delta$ Acc. (pp) | 95% CI | Raw $p$ | Holm $p$ |
|---|---|---|---|---|
| Static graph only | +5.5 | [+4, +7] | $< 0.0001$ | $< 0.0001$ |
| Dynamic graph only | +2.2 | [+0.9, +3.5] | 0.0008 | **0.0024** |
| No semantic rerank; links retained | +3.1 | [+1.8, +4.4] | 0.0002 | **0.0012** |
| No confidence rerank | +2.6 | [+1.3, +3.9] | 0.0004 | **0.0020** |
| Dense retrieval; no links | +4.5 | [+3.2, +5.8] | $< 0.0001$ | $< 0.0008$ |
| Raw verifier confidence | +2 | [+0.7, +3.3] | 0.0020 | **0.0024** |
| No source-auth gate | +2.9 | [+1.5, +4.3] | 0.0004 | **0.0020** |
| No conflict isolation | +2.5 | [+1.1, +3.9] | 0.0008 | **0.0024** |

# I Matched Continual-memory and Uncertainty Controls

MedQA editors share feedback, generator, caps, windows, and initial memory but evolve private states. HippoRAG 2 ingests revealed records; graph editors share frozen extraction/linking candidates. Post-commit Qwen extraction, MedGemma scoring, admission, and publication to first visibility are outside answer cost. Dataset keys supervise updates; audit labels remain withheld. Confidence-aware KG is paper-specified. Conflict precision $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ uses native pre-evaluation flags and fixed authorized-source Selene labels, blinded to verifier/method/admission and unchanged by detector ablations.

**Paired update effects on MedQA.** For immediate calibrated admission versus no update, the 25 paired trajectory differences have AULC mean $\bar{d} = 5.30$ percentage points, sample standard deviation $s_d \approx 0.61$ points, and $T_{\mathrm{obs}} \approx 43.4$. With $b = 0$ extreme draws among 100,000, the primary wild-test

Table 21: Effects on 25 paired trajectories. A: MedQA immediate minus no update. B: MedMCQA schedules, with separate AULC/Final effects and CIs. CIs and raw AULC $p$ values are unadjusted; B's Final has intervals only.

**A. MedQA:** AULC $\Delta = +5.3$ (95% CI $[+5.06, +5.54]$); Final $\Delta = +5.4$ (95% CI $[+5.24, +5.56]$).

**B. MedMCQA**

| Contrast | Δ AULC | AULC 95% CI | Δ Final | Final 95% CI | $p_{\mathrm{AULC}}$ |
|---|---|---|---|---|---|
| Immediate − No update | +3.6 | [+2.8, +4.4] | +3.7 | [+2.9, +4.5] | 0.00001 |
| CM immediate − No update | +3.2 | [+2.4, +4.0] | +3.1 | [+2.3, +3.9] | 0.00001 |
| CM delayed − No update | +1.4 | [+0.6, +2.2] | +1.4 | [+0.5, +2.3] | 0.002 |
| CM permuted − No update | +1.7 | [+0.9, +2.5] | +1.9 | [+1.1, +2.7] | 0.00001 |
| Immediate − CM immediate | +0.4 | [+0.1, +0.7] | +0.6 | [+0.2, +1.0] | 0.021 |
| CM immediate − CM delayed | +1.8 | [+1.1, +2.5] | +1.7 | [+1.0, +2.4] | 0.00001 |
| CM immediate − CM permuted | +1.5 | [+0.8, +2.2] | +1.2 | [+0.5, +1.9] | 0.00001 |

Differences are percentage points. MedMCQA immediate minus no update AULC: $b = 0$, $B = 100{,}000$, $p = 1/100001$; other $p$ values use the displayed precision.

probability is $p = 1/100001$. For Final, $\bar{d} = 5.40$ points and $s_d \approx 0.41$ points. Paired confidence intervals appear in Table 21.

### I.1 Supervised DC-Cu: Paired Effects and Curation Contract

Supervised DC-Cu adapts Suzgun et al. (2026)'s text curation with matched MedQA post-commit feedback/budgets across 25 paired trajectories (Table 22).

**Curation and memory updates.** After commitment, Qwen3.5-9B curates the complete feedback package (question, released options, true answer) into a reusable strategy fragment. Split curated text into segments of at most 256 wordpieces with 32-token overlap; drop segments under eight wordpieces. Before insertion, identical segments merge provenance; near-duplicates with cosine similarity $\geq 0.95$ merge into the longer segment. There is no conflict quarantine.

**Retrieval.** Index the deduplicated strategy segments using `all-mpnet-base-v2` and FP32 $\ell_2$ normalization. Exact cosine search returns the top 20 segments within 1,024 tokens, breaking ties by record ID.

**Curation prompt.** Qwen3.5-9B uses the following instruction; `[PACKAGE]` is the completed question, options, and released true answer text.

**System.** Curate the completed question and released true answer into a compact, reusable strategy fragment. Preserve only what is supported by the completed question and the released answer. Do not copy raw questions, option letters, or item-ID lookups. Do not treat the model's predicted answer as truth. Return plain text only.
**User.** COMPLETED QUESTION AND TRUE ANSWER: `[PACKAGE]`

Table 22: MedQA: EvoMedAgent minus Supervised DC-Cu. Effects/SDs/CIs are in percentage points; tests are unadjusted. Sample SD uses $\mathrm{ddof} = 1$ across 25 paired trajectories.

| Endpoint | Mean difference | Sample SD | Population SD | 95% CI | Raw $p$ | Positive |
|---|---|---|---|---|---|---|
| AULC | +1.24 | 0.62 | 0.6075 | $[+0.99, +1.49]$ | 0.00001 | 24/25 |
| Final | +1.08 | 0.5715 | 0.5600 | $[+0.88, +1.3]$ | 0.00001 | 25/25 |

Population SD ($\mathrm{ddof} = 0$) equals sample SD$\times\sqrt{24/25}$; AULC uses the rounded sample SD.

## J Graph-state Contract and Runtime Accounting

Frozen evaluation preserves initial memory, with read-only EHR patient views. MedQA/MedMCQA online and EHR upstream phases apply accepted events after commitment, scoring, and reveal; EHR downstream preserves the boundary snapshot.

Answer-path p50/p95 is 1.28/2.03 seconds; Table 33's post-commit stage sums are not total-latency quantiles. After 100 untimed warm-ups, full-stream timing uses a monotonic clock and boundary

Table 23: Matched curation: Qwen3.5-9B answers, dataset keys score, Selene Mini labels offline evidence. The shared cap covers retrieved context.

| Component | EvoMedAgent | Supervised DC-Cu |
|---|---|---|
| Curator | Qwen3.5-9B extraction; MedGemma verification | Qwen3.5-9B curation |
| Feedback | Post-commit dataset answer | Post-commit dataset answer |
| Write unit | Scoped assertion | Text strategy fragment |
| Retrieved evidence cap | $K = 20$, $K_d = 5$, 1,024 tokens | Top-20 passages, 1,024 tokens |
| Conflict handling | Quarantine of unresolved conflicts | No conflict quarantine |
| Source calibration | Cross-model source-support calibration | None |

Table 24: Serialized-context distributions for Deep-routed questions after all edge and token caps.

| Dataset | Tokens p50 | Tokens p95 | Tokens max |
|---|---|---|---|
| MedQA | 322 | 387 | 505 |
| MedMCQA | 437 | 481 | 486 |
| PubMedQA | 458 | 496 | 540 |
| MMLU-Gene | 328 | 418 | 546 |
| MMLU-Clinic | 484 | 512 | 524 |
| EHRNoteQA | 338 | 396 | 538 |

CUDA synchronization. The shared A800 host uses loopback, 16 physical cores, 128 GiB cgroup RAM, one OpenMP/MKL/FAISS thread, fixed affinity, and performance governor. FAISS 1.14.3 follows Table J.2; Neo4j Community 5.26.30 uses one local database/Bolt connection, 16 GiB heap, and 64 GiB page cache. Read-only answers and updates are timed separately.

Table J.2: FAISS CPU index. MedCPT article `[CLS]` vectors use float32 and L2 normalization before indexing. $M$ is neighbors per layer; `efConstruction`/`efSearch` are build/query search widths.

| Index type | $M$ | `efConstruction` | `efSearch` | Dimension | Metric | Normalization |
|---|---|---|---|---|---|---|
| `IndexHNSWFlat` | 32 | 200 | 128 | 768 | Inner product | L2 |

## K System Scores Across Backbones

Table 25 pins six open-weight and three hosted backbones; tokenizers/templates share model revisions and roles follow Appendix D. The offline judge is `AtlaAI/Selene-1-Mini-Llama-3.1-8B`. Hosted endpoints `gpt-5.6-luna`, `gpt-5.6-terra`, and `gpt-5.6-sol` use density-only routing and text-only inputs.

Table 25: Nine-backbone execution configuration. Open-weight capacities are vendor-reported; gpt-oss gives nominal/active parameters.

| Backbone (linked revision) | Access | Capacity | Execution | Route |
|---|---|---|---|---|
| Qwen3.5-0.8B | local | 0.8B | A800; BF16 | H+D |
| MedGemma-27B-Text-IT | local | 27B | A800; BF16 | H+D |
| Ministral 3 8B | local | 8.4B | A800; BF16 | H+D |
| Qwen3.5-9B | local | 9B | A800; BF16 | H+D |
| gpt-oss-20B | local | 21/3.6B | A800; MXFP4→BF16 | H+D |
| Qwen3.8-27B-FP8 | local | 27B | A800; FP8→BF16 | H+D |
| GPT-5.6 Luna | API | cost-sensitive | Responses API | D |
| GPT-5.6 Terra | API | balanced | Responses API | D |
| GPT-5.6 Sol | API | flagship | Responses API | D |

**Software and request execution.** The environment specification is Python 3.11.13, PyTorch 2.8.0+cu128, Transformers 5.16.1, Accelerate 1.10.1, `huggingface-hub` 1.5.0, `tokenizers` 0.23.1, `safetensors` 0.8.0, NumPy 2.2.6, SciPy 1.15.3, and scikit-learn 1.7.2. Release versions and core dependencies are pinned to official metadata.

Local 0.8B–27B configurations use one A800 80GB, `torch.inference_mode()`, eager attention, BF16 activations/accumulation, no TF32, and no offload. Qwen3.8 FP8 and gpt-oss-20B MXFP4 weights expand to BF16; Qwen3.5, MedGemma, and Ministral load native BF16. Qwen disables thinking; Qwen3.8 also disables preserved thinking history. gpt-oss-20B uses Harmony with reasoning effort low.

Local answers use repetition penalty 1.0, batch size 1, a 4,096-token input cap, and Appendix D's decoding. Each request records its dataset, item, method, backbone, run identifier, and seed. Methods share order and replicate indices but use method-specific sampling streams; statistical seeds are separate. Probes use FP32 stable log-softmax with padding masked; other calls follow Appendix D.

Hosted Responses API calls use low reasoning effort/verbosity, no tools, disabled request storage, the same 256/512 output caps, and supported fields only. Direct, Always-Deep, and EvoMedAgent-D calls are interleaved per item, each repeated five times and scored separately. API latency includes network time and is reported separately from local-GPU latency.

Table 26: **Nine-backbone system accuracy (%).** Local policies use H+D; hosted policies use EvoMedAgent-D. Policies are selected before test; five answers per fixed item are scored separately and averaged. MMLU-Gene has $N = 100$, $R = 5$ (500 answers).

| Backbone | Route | MedQA | MedMCQA | PubMedQA | MMLU-Gene | MMLU-Clinic | EHRNoteQA |
|---|---|---|---|---|---|---|---|
| Qwen3.5-0.8B | H+D | 62.8 | 59 | 72.2 | 67.8 | 69 | 56.4 |
| MedGemma 27B | H+D | 83.5 | 65.7 | 74.4 | 74 | 77.2 | 83.6 |
| Ministral 3 8B | H+D | 71.9 | 68.7 | 79.8 | 75.6 | 76.4 | 63.8 |
| Qwen3.5-9B | H+D | 82.1 | 77 | 86.6 | 92 | 91.7 | 78.4 |
| gpt-oss-20B | H+D | 72.5 | 69.1 | 80.4 | 79 | 80.2 | 66.2 |
| Qwen3.8-27B-FP8 | H+D | 78.1 | 73.9 | 84.6 | 87 | 87.8 | 71.5 |
| GPT-5.6 Luna | D | 92.9 | 80.7 | 90.4 | 95 | 94.8 | 78.8 |
| GPT-5.6 Terra | D | 94 | 83 | 92 | 96.8 | 96 | 81.1 |
| GPT-5.6 Sol | D | 96.2 | 84.6 | 93.2 | 98 | 97.4 | 83 |

MMLU-Gene: MedGemma, Qwen3.8, and Luna have 370, 435, and 475 correct answers out of 500, respectively.

## L Extraction Audit and Noise Injection

Interventions test unauthorized sources, corrupted proposals, redirected endpoints, reversed polarity, and new-ID duplicates; semantic attacks retain source spans. Module tests measure admission/visibility under authorization, schema/link, calibration, conflict, and rate/source-mixture checks.

**Source-derived references.** Source audits use 1000 linking references, 1000 triple references, and 1000 static edges.

Provenance/patient/article/near-duplicate families follow Appendix D.1. Before inspecting model outputs, known-probability sampling stratifies dataset, source, relation, negation, time, entity ambiguity, and length. Frozen Selene inventories and independent top-32 UMLS/SapBERT mapping follow Appendix D. One-to-one matching requires mention-span overlap plus CUI, or canonical subject/relation/object, polarity, interval, and support span. Byte offsets, CUI, and triple keys break ties; duplicates match once. Inclusion-weighted micro precision/recall divide matches by predictions/eligible references. Table 27's inventory $N$ counts reference objects for linking/triples and audited static edges. Link validation, static fitting, and final audits are family-disjoint.

**Static reliability.** Using only fit families and seed 62001, sample without replacement up to 200 edges per frozen source-class–relation stratum (all edges if smaller). The blinded assertion prompt (Appendix D) yields Yes/No/Unresolved without dynamic verifier outputs. Set $r_s = (S_s + 0.5)/(n_s + 1)$, where $S_s$ counts Yes and $n_s$ includes No/Unresolved; empty strata score zero. This Jeffreys-form smoothed ranking score freezes before policy selection. Disjoint report families evaluate resolved-label ECE/Brier, coverage, unresolved score distributions, and all-zero/all-one label bounds.

Table 27: Selene-reference linking, extraction, and static support. Inventory and report $N/N_+$ give resolved/positive label counts in separate populations. Precision/recall use source inventories; ECE/Brier use the blinded report partition.

| Stage | Inventory $N/N_+$ | Report $N/N_+$ | Proxy precision | Proxy recall | ECE | Brier |
|---|---|---|---|---|---|---|
| Entity linking | 1000/868 | 240/208 | 0.958 | 0.868 | – | – |
| Triple faithfulness | 1000/914 | 220/201 | 0.927 | 0.914 | 0.028 | 0.076 |
| Static-edge support | 1000/912 | 260/237 | 0.884 | – | 0.037 | 0.078 |

Table 28: Selene-reference calibration and gate diagnostics. $E$ counts reference-negative score/recheck passes before source–relation activation; $E/N$ uses all resolved report candidates. Dash: closed `causes` group; its 90 candidates remain in total $N$.

| Stratum | Report $N/N_+/N_-$ | Raw ECE | Calibrated ECE | Precision lower bound | Full coverage | $E$ | Pre-activation error yield $E/N$ |
|---|---|---|---|---|---|---|---|
| `has_finding` | 310/279/31 | 0.036 | 0.049 | 0.932 | 0.9 | 17 | 0.055 |
| `associated_with` | 190/172/18 | 0.055 | 0.051 | 0.91 | 0.7 | 10 | 0.053 |
| `indicated_for` | 165/156/9 | 0.052 | 0.032 | 0.942 | 0.8 | 8 | 0.048 |
| `contraindicated_for` | 120/111/9 | 0.058 | 0.033 | 0.92 | 0.8 | 2 | 0.017 |
| `treated_with` | 210/193/17 | 0.041 | 0.029 | 0.912 | 0.8 | 13 | 0.062 |
| `measured_by` | 140/127/13 | 0.046 | 0.059 | 0.918 | 0.9 | 10 | 0.071 |
| `causes` | 90/80/10 | 0.037 | 0.027 | – | 0 | 0 | – |
| **Relation total** | **1,225/1,118/107** | – | – | – | – | **60** | **0.049** |
| Source: dataset-linked audit | 480/442/38 | 0.022 | 0.034 | 0.92 | 0.7 | 24 | 0.050 |
| Source: authorized note | 320/285/35 | 0.039 | 0.032 | 0.89 | 0.6 | 18 | 0.056 |
| Source: guideline | 425/391/34 | 0.057 | 0.035 | 0.93 | 0.6 | 18 | 0.042 |
| **Source total** | **1,225/1,118/107** | – | – | – | – | **60** | **0.049** |

**Which candidates each calibration metric uses.** Table 28 partitions 1,225 resolved report candidates (1,118 positive, 107 negative; equal-weight Brier 0.0755) by relation and source; the partitions are not additive. Source-audit stages use separate populations (Table 27).

Full coverage divides complete-gate admissions by write-eligible candidates. $E$ counts negative-labeled score/recheck passes before source–relation activation; $E/N$ is their fraction of all resolved candidates, including rejections. It does not measure errors admitted to memory. Both partitions give $60/1{,}225 = 0.049$ (rounded).

**Evaluation populations.** Each population records its ID, members, unit (candidate, canonical edge, query–edge, or query), scored/resolved/positive/negative/unresolved counts, positive fraction, target, raw/calibrated/static-source probabilities, weights and their source, and exclusions. Table 27's inventory $n = 1000$ is separate from ECE/Brier observations. For equal-weight resolved candidates,

$$\mathrm{ECE} = \sum_{b=1}^{15} \frac{|S_b|}{N} \left|\overline{p}_b - \overline{y}_b\right|, \qquad \mathrm{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2. \tag{17}$$

Empty bins contribute zero. Precision/recall inclusion weights do not carry over to ECE, whose weighting is defined separately. Source support and joint admission are distinct targets, unchanged by disabling a gate. Raw/calibrated scores share 15 equal-width bin edges, not necessarily bin membership.

**No-conflict candidate bridge.** $S_{\text{No-conflict}} = S_{\text{Full}} \dot{\cup} A$, $|A| = 196$. Candidate/assertion/case IDs link roles, conflict/support/joint labels, and No-conflict membership; shared candidates retain labels, probabilities, and weights. Cases may contain multiple candidates; unresolved conflict need not imply unresolved source support.

**Pooled retrieval population.** For paired queries $Q$, the shared relevant pools are $G_q = \{e : \mathrm{relevance}(q, e) \in \{1, 2\}\}$. Each method is evaluated as

$$R@20_q^m = \begin{cases} |R_q^m \cap G_q|/|G_q|, & |G_q| > 0, \\ 0, & |G_q| = 0, \end{cases} \qquad R@20^m = \frac{1}{|Q|}\sum_{q\in Q} R@20_q^m. \tag{18}$$

MedQA has 312 queries, 18 empty pools (Table H.2). $R_q^m$ uses `ranked_top20` before packing, versus top-100 pool construction (Appendix H). "Pooled" merges reference pools, not query counts; shorter lists leave $G_q$ unchanged. Unresolved labels are excluded from ECE/Brier and included in

Table L.2: Resolved candidate populations for Write ECE/Brier. Positive and negative counts use the joint evaluation target.

| Population | $N_{\text{resolved}}$ | $N_+$ | $N_-$ | Positive fraction |
|---|---|---|---|---|
| Full | 8,921 | 5,120 | 3,801 | 0.574 |
| Static only | 5,913 | 3,300 | 2,613 | 0.558 |
| Dynamic only | 3,008 | 1,820 | 1,188 | 0.605 |
| Raw confidence | 8,921 | 5,120 | 3,801 | 0.574 |
| No source-auth gate | 8,921 | 5,120 | 3,801 | 0.574 |
| No conflict | 9,117 | 5,240 | 3,877 | 0.575 |

Static and dynamic subsets sum to Full. The disjoint extension $S_{\text{No-conflict}} = S_{\text{Full}} \dot{\cup} A$, $|A| = 196$, adds 120 positive and 76 negative joint-target labels. The counts 148/30/18 in Table 6 refer to conflict labels, whereas this table uses the joint evaluation target. Positive fractions are rounded to three decimals.

the support-rate denominator. Table D.1 gives the quality-partition $A_s, X_s$ counts; $A_s - X_s$ includes Unresolved reference judgments.

Read ECE uses resolved retrieved edges. Bins are left-closed/right-open, with one included in the last bin. Marginal support rates count unresolved references as nonsuccess; report bounds accompany their support estimates. nDCG@20 uses gain $2^{\text{grade}} - 1$, discount $\log_2(\text{rank} + 1)$, and the highest 20 pool grades for IDCG. Zero-IDCG queries score zero and are counted separately. Queries have equal weight. Selection and reference prompts remain fixed before test comparisons.

**Extracted-assertion interventions.** Use 500 of the 1,273 MedQA items routed Deep with dynamic evidence under clean EvoMedAgent. Copy each history, disable feedback updates, and retain H+D, $K = 20$, $K_d = 5$, 1,024 graph tokens, Qwen3.5-9B generation/extraction, MedGemma-27B verification, and full ADMIT. Each item receives one greedy answer per replicate (`do_sample=false`). Object replacement/polarity reversal acts before linking; endpoint redirection acts before verification. Source case and support span remain fixed, while perturbed assertions receive identity $\kappa(e')$; pairing case IDs are separate. Verification, calibration, schema, authorization, and conflict checks rerun before copied transactions reduce and QA begins. False accept divides retrieval-eligible unsupported injections by all model-rated unsupported injections. Five corruption seeds share clean IDs, surface-frequency/source strata, and items; accuracy averages their evaluations. Each clean greedy replicate gives 378/500 (75.6%); the 82.1% frozen result instead averages five stochastic answers on all 1,273 items (Table 1).

Table 29: Copied-state assertion corruption. False accept: unsupported injections passing the complete write gate; clean "–": empty injection denominator.

| | 10% corruption | | 20% corruption | |
|---|---|---|---|---|
| Corruption | Accuracy | False accept | Accuracy | False accept |
| Type-compatible object | 74.3 | 0.058 | 72.9 | 0.079 |
| Nearest-neighbor endpoint | 73.9 | 0.041 | 73.2 | 0.071 |
| Polarity flip | 74.6 | 0.049 | 72.7 | 0.066 |
| Equal-probability mixture | 73.7 | 0.077 | 72.4 | 0.093 |
| Clean (0% corruption): accuracy 75.6; false accept –. | | | | |

PubMedQA uses 500 items with five answers each; Table 1's descriptive gains over Always-Deep/HippoRAG 2 are 1.2/1.7 points.

Table 30: Frozen paired effects (EvoMedAgent minus comparator); $N$ counts five-answer item/patient blocks. Inference is unadjusted (Appendix A); MedQA HippoRAG 2 shares scores/CI/raw $p$ with Table 19.

| Dataset | Comparator | Paired $N$ | $\Delta$ Acc. (pp) | 95% CI | Raw $p$ |
|---|---|---|---|---|---|
| MedQA | Always-Deep | 1273 | +1.6 | [+0.3, +2.7] | 0.011 |
| MedQA | HippoRAG 2 | 1273 | +2 | [+0.8, +3.4] | 0.0021 |
| MedMCQA | Always-Deep | 4180 | +2 | [+0.7, +3.1] | 0.0075 |
| MedMCQA | HippoRAG 2 | 4180 | +2.4 | [+1.2, +3.8] | 0.0006 |
| EHRNoteQA | Always-Deep | 500 | +1.8 | [+0.4, +2.8] | 0.015 |
| EHRNoteQA | HippoRAG 2 | 500 | +2.1 | [+0.8, +3.6] | 0.0021 |

# M Additional Online and Backbone Comparisons

## M.1 Comparison with AMG-RAG and ACE

Table 31 compares published AMG-RAG (Rezaei et al., 2025) and ACE (Zhang et al., 2026) mechanisms. Answer-before-update is established; our contribution concerns medical scope checks, source-support calibration, and joint admission.

Table 31: Comparison of published memory-policy specifications.

| Dimension | EvoMedAgent | AMG-RAG | ACE |
|---|---|---|---|
| Feedback | Post-commit dataset answers | Retrieved medical sources; query-driven expansion | Execution feedback; optional reference labels |
| Write unit | Scoped assertion | Medical entity/relation with metadata | Structured playbook bullet and delta |
| Support scoring | Cross-model source-support calibration | LLM relation confidence; path propagation | Helpful/harmful counters |
| Conflict handling | Quarantine of unresolved conflicts | | Corrective edits and pruning |
| Patient scope | Patient identity; physiological, clinical, medication/intervention scope | Clinical context | |
| Retrieval budget | $K = 20$, $K_d = 5$, $B = 1024$ tokens | Native threshold/document cap | Native playbook/context budget |

Table 32: Online endpoints and CM terminal counts/trajectory; endpoints average 25 trajectories. CM restores initial memory during positions 1–100 and varies feedback release.

| Dataset | Policy | Final | AULC | Transactions/run | Edges/run | First-100 AULC | Post-burn-in AULC |
|---|---|---|---|---|---|---|---|
| MedQA | CM immediate | 87.1 | 86.5 | 6082 | 2147 | 70.7 | 88.26 |
| MedQA | CM delayed | 86.1 | 85.5 | 6082 | 2147 | 70 | 87.22 |
| MedQA | CM history-permuted | 86.3 | 85.7 | 6082 | 2147 | 70 | 87.44 |
| MedMCQA | No update | 76.5 | 76.2 | – | – | 76.1 | 76.2 |
| MedMCQA | Immediate | 80.2 | 79.8 | – | – | 76.8 | 79.9 |
| MedMCQA | CM immediate | 79.6 | 79.4 | 8213 | 1406 | 65.0 | 79.8 |
| MedMCQA | CM delayed | 77.9 | 77.6 | 8213 | 1406 | 65.0 | 78.0 |
| MedMCQA | CM history-permuted | 78.4 | 77.9 | 8213 | 1406 | 65.0 | 78.3 |

For MedQA $T = 1{,}000$ and MedMCQA $T = 3{,}500$, $A_{\text{all}} = (100/T)A_{1:100} + ((T - 100)/T)A_{101:T}$ before rounding; MedQA post-burn-in values retain four decimals before display rounding. Counts represent visible ADMIT transactions/merged dynamic assertions. CM contrasts jointly vary available evidence (including masked initial memory) and update timing.

Table 33: Stage costs. The 27.7% saving covers answering; post-answer totals sum stage quantiles arithmetically.

| Stage | p50 s | p95 s | Answer path |
|---|---|---|---|
| Always-Deep answer path | 1.77 | 2.86 | Yes |
| EvoMedAgent answer path | 1.28 | 2.03 | Yes |
| Online extraction | 0.42 | 0.88 | No |
| Online verification | 0.18 | 0.41 | No |
| Publication/indexing | 0.09 | 0.21 | No |
| Summary | Sum of p50 s | Sum of p95 s | Scope |
| Post-answer stages | 0.69 | 1.50 | After commitment |

**Feedback availability and corruption.** Conditions share 1,000 MedQA items, 25 trajectories, one answer/item, and generator/decoding/budgets/initialization. Figure 5 gives full-pool endpoints; complete feedback uses immediate calibrated updates.

PCG64 samples independent Bernoulli($p$) visibility, $p \in \{0.10, 0.25, 0.50\}$, fixed before prediction and shared across methods within trajectories and repeated item generations. Feedback follows commitment/scoring; admissions appear next item. Unselected items produce no feedback/proposals and enter neither corruption masks nor counts. All gates remain active.

Available feedback is independently corrupted at $q \in \{0.10, 0.30\}$; standalone noise releases all feedback. Shared pre-prediction masks/replacements use PCG64 uniform non-gold-option sampling in released order (binary: the other option), retaining true/wrong indices/texts. Only `correct_answer_text` changes; questions, option order, item IDs, and order/generation seeds stay fixed. Original truth scores predictions; admission counts corrupted-feedback candidates, separating source entailment from answer correctness.

Each corruption condition retains its extraction view (stem and replacement answer, no options). Spans/byte offsets are recomputed and checked against this view. Original source ID/version identify the upstream source; class, provenance, patient ID, observation time, commit record, and authorization remain fixed. MedGemma, calibration, and source/scope/conflict gates rerun. This tests corrupted feedback robustness, not alteration detection against the unmodified source. Table 34 separates items from the assertion cascade; the score gate is $\widehat{r} \geq 0.70$.

Table 34: Feedback masks and assertion cascade: mean counts over 25 trajectories of 1,000 items each, displayed as integers. Sparse/noise masks count feedback-visible/replaced items; later columns count gate-passing assertions. All rows use the source checks above; noise-row source rejections reflect pre-existing defects.

| Condition | $p/q$ | Masked items | Extracted assertions | Span pass | Calibrated $\geq 0.70$ | Source pass | Admitted | Retrievable |
|---|---|---|---|---|---|---|---|---|
| sparse-10 | 0.10 | 102 | 186 | 181 | 92 | 88 | 83 | 83 |
| sparse-25 | 0.25 | 247 | 449 | 437 | 218 | 209 | 197 | 197 |
| sparse-50 | 0.50 | 503 | 915 | 890 | 443 | 425 | 401 | 401 |
| noise-10 | 0.10 | 103 | 187 | 182 | 88 | 84 | 79 | 79 |
| noise-30 | 0.30 | 298 | 542 | 521 | 247 | 231 | 218 | 218 |

Mean source rejections per trajectory are 4/16 in noise-10/noise-30, comprising pre-existing unversioned/invalid-version sources (1/3), incomplete provenance (2/11), and patient-binding inconsistency (1/2): missing observation time, commit records, and malformed source IDs contribute 1/5, 1/4, and 0/2, respectively. None concerns unauthorized families or generated answers. Unchanged metadata yields identical failures, counted separately from corruption-induced semantic failures.

**Selection and domain conditions.** Uncertainty selection fixes the top 250 pre-run entropy items (Appendix D). Cardiovascular ($D_1$), infectious disease ($D_2$), neurology ($D_3$), and endocrinology ($D_4$) each contain 250 items; shifted order is $D_1$–$D_2$, then $D_3$–$D_4$. Domain-selected feedback releases the first 500 answers only and scores the last 500, restarting AULC at $T = 500$, $W = 100$, with last-100 Final: 85.9/86.6. Other rows use $T = 1{,}000$, last-200 Final; domain controls match order/scoring population. One logical hour equals ten positions. The 24-hour delay publishes $t$ feedback after $t + 239$, readable at $t + 240$; post-stream arrivals cannot affect scores.

## N Patient-disjoint Cross-stream Memory Transfer

Each of 25 histories independently permutes 462 update and 500 disjoint EHRNoteQA evaluation patients (Kweon et al., 2024), starting from PrimeKG and empty dynamic memory. Transfer and masking use Always-Deep, Qwen3.5-9B, temperature 0.7, top-$p$ 0.95, $K = 20$, $K_d = 5$, $B = 1{,}024$. Transferred null-ID assertions undergo source/ten-field scope checks; the enabled note relation is `has_finding` (Table D.1). J0 uses Appendix D's masked Selene rubric/adapter. J1 is pinned Claude Sonnet 4.6 (`claude-sonnet-4-6`) (Anthropic, 2026), rescoring EHR answers once at temperature 0, thinking off, 512-output-token cap. Both use 1–5 scores, $\geq 4$ correct. J1 receives

question/reference/original-record-view/candidate; method, masking, rank, and support scores are hidden.

Table 35: EHR common-set masking. A: trajectory–patient requests, five answers each. B/C: 25 equally weighted trajectories; random repeats average within trajectory. C: marginal paired 95% intervals (percentage points).

**A. Exposure and evaluation**

| Stage | Requests / denominator | Coverage (%) |
|---|---|---|
| Target present / all requests | 3,600 / 12,500 | 28.80 |
| Matched / target present | 2,400 / 3,600 | 66.67 |
| Matched / all requests | 2,400 / 12,500 | 19.20 |
| Reference exclusions / matched | 24 / 2,400 | 1.00 |
| Common evaluable / all requests | 2,376 / 12,500 | 19.008 |

**B. Accuracy and pooled counts**

| Method | J0 mean (%) | J1 mean (%) | J0 pooled $C/N$ | J1 pooled $C/N$ |
|---|---|---|---|---|
| Full | **77.2277** | **75.9141** | 9,167/11,880 | 9,016/11,880 |
| Target masking | 74.3997 | 73.0710 | 8,831/11,880 | 8,675/11,880 |
| Random masking | 76.6563 | 75.7273 | 90,980/118,800 | 89,884/118,800 |

**C. Paired effects**

| Contrast | J0 difference | J0 95% CI | J1 difference | J1 95% CI |
|---|---|---|---|---|
| Full − Target | +2.8280 | $[2.4757, 3.1783]$ | +2.8431 | $[2.3490, 3.3244]$ |
| Full − Random | +0.5714 | $[0.3489, 0.7783]$ | +0.1868 | $[-0.3277, 0.6945]$ |
| Random − Target | +2.2566 | $[1.9063, 2.6207]$ | +2.6563 | $[2.1124, 3.1903]$ |

Pooled $C/N$ are checksums; random counts weight reused answers, not independent generations. Effects condition on matchable, evaluable requests. The J1 Full–Random interval includes zero.

**Evaluation population.** Uniformly excluding five reference-insufficient patients across trajectories, branches, and judges leaves $25 \times 495 = 12{,}375$ requests (61,875 answers), Full J0/J1 74.2917/73.0925%. Their 24 matched requests remove 120 answers; Table 35 gives the common set. Of successful matches, 1,800/600 have one target and one/two controls; 900 failures lack controls and 300 have both dynamic assertions targeted. Matched/common-evaluable requests cover 480/475 distinct patients, the latter 475/500 (95%). Transfer inference resamples 25 paired trajectories after within-trajectory mask averaging. Accuracy averages $100C_h^m/N_h$; unequal $N_h$ separates it from pooled $C/N$. Missing/unparseable answers count as errors after shared reference exclusions.

**Target and matched random masking.** Annotators A/B, blinded to reference/generated answers and scores, adjudicate and freeze Full-context targets: upstream-learned null-ID `has_finding` assertions with complete provenance, query-matching entities/relations, and patient-compatible conditions. Same-context non-target controls undergo the same complete-provenance check, then match without replacement on upstream null-ID origin, source class, relation, deletion count, token length within 10%, and dynamic position within one slot; ties are random. Seeds 91001–91010 give ten masks, retaining IDs/pairings/token counts. Identical within-request deletion sets reuse five answers, weighted by mask multiplicity. No refill, retrieval, or static substitution occurs; other content/order stays fixed. All-target/control-free contexts are unmatchable. Update patients' five answers share pre-feedback memory and commit before publication; downstream memory is frozen.

**Transfer baselines.** Method-specific Acc./Final: No update 71.0/72.0, Full 74.5/76.0, KEDKG 73.0/74.0, HippoRAG 2 72.5/74.0, supervised DC 71.5/72.0. Full–No update: +3.5 points (95% CI $[2.8, 4.2]$); common-evaluable/exclusions-as-wrong CIs: $[2.7, 4.2]$/$[2.8, 4.2]$. These baseline populations differ from the masking common set.

### N.1 Transfer chains and source audit

**Batch-II sampling and review.** Within all 25 trajectories, Batch II samples without replacement from 2,400 matched pre-exclusion Full requests (480 patients), retaining five answers each. Seed 93001 draws 100 of the 7,200 entries without replacement. J0's pinned Selene Mini reuses Appendix D's prompt. `completed_question`/`correct_answer_text` contain the unchanged upstream EHRNoteQA question/released feedback from the assertion's

first-publication source (primary source after merging). Missing answers stay empty, add no conditions, and are never filled from downstream answers or predictions. The de-identified `source_span` and `canonical_assertion` with stored ten-field `scope` complete the input; method/rank/support/admission are hidden. Support also requires rule-based patient compatibility (Appendix H.1) and assertion-ID presence in downstream context logs; neither downstream context nor current patient state enters the model prompt. Human checks cover source location, direction/negation, population level, and condition completeness. Review covers eight model-unresolved/automatically unlocatable entries, with labels visible. Review changes two supported labels to unresolved and one erroneous label to supported: 93/6/1 becomes 92/5/3 (supported/error/unresolved). All 100 remain in the denominator; this is not independent clinical validation. The three unverifiable sources trace to parser-exception admission in this sample.

Table N.4: Transfer counts and Batch-II audit across 25 trajectories.

**A. Transfer-stage counts**

| Unit | Trajectory batch I Sum | Trajectory batch I Unique | Trajectory batch II Sum | Trajectory batch II Unique |
|---|---|---|---|---|
| Accepted candidate events | 1,842 | 1,210 | 1,936 | 1,247 |
| Published null-ID assertions | 1,210 | 876 | 1,247 | 907 |
| Valid snapshot assertions | 1,185 | 852 | 1,209 | 881 |
| Assertions retrieved downstream | 876 | 640 | 918 | 653 |
| Answer-context entries | 3,420 | – | 7,200 | – |
| Downstream patients | – | 498 | – | 480 |

**B. Chain audit**

| Check | Numerator/denominator | % |
|---|---|---|
| Complete snapshot links | 1,171/1,209 | 96.86 |
| Context linked to source | 7,083/7,200 | 98.38 |
| Context linked to answer ID | 7,126/7,200 | 98.97 |
| Context-subset patients | 480/500 | 96.00 |
| Sample source-checkable | 97/100 | 97.00 |
| Population support and conditions | 92/100 | 92.00 |
| Definite generalization/scope error | 5/100 | 5.00 |
| Unresolved source verification | 3/100 | 3.00 |

**C. Sample outcomes**

| Outcome | Count | % |
|---|---|---|
| Supported, conditions intact/applicable | 92 | 92.00 |
| ID-removal generalization | 2 | 2.00 |
| Condition omission/scope expansion | 3 | 3.00 |
| Incomplete source version | 2 | 2.00 |
| Source-span location failure | 1 | 1.00 |
| Total | 100 | 100.00 |

Sum aggregates; Unique deduplicates assertion/patient IDs. Batches have different seeds. I filters all requests for provenance/span/answer-link completeness. II's 918/653 cover all requests; 7,200/480 cover its pre-exclusion matched-request sample without that filter. Context keys: trajectory/patient/answer/assertion, retaining repetitions without mask weights. Linkage rates are marginal; candidates precede deduplication.