Evaluating improvement in healthcare agents

Design and reporting templates accompanying RSI-HA | Resource version 0.3.0 | Search cutoff 4 November 2026

Choose the question before choosing the control

Memory contribution, generator contribution and a better improvement procedure are different questions. These templates preserve the review's existing concepts and task-native outcomes. They do not define a new score or a validated reporting standard. Verification status of the underlying extraction is documented in REVIEW_STATUS.md.

1. Does persistent memory help?

Comparison. Paired persistent-memory condition versus memory-disabled/empty condition. A frozen prebuilt-memory arm answers a different question from ongoing writeback.

Keep fixed or disclose. Task backbone, orchestration, tools, starting policy and base artifacts, feedback access and initial cases. Keep the patient/case sequence paired or record the ordering difference.

Timing and independent tests. Write back only after the current case. State which labels become available and when. Distinguish an online adaptation stream from an untouched final test.

2. Does a changed generator or recipe help?

Comparison. Changed generator/operator/skill/recipe versus the corresponding frozen version, followed by matched learner training and independent downstream evaluation.

Keep fixed or disclose. Shared learner start, task model, training protocol, data access, selection rule and downstream evaluator. State whether accepted-data volume, learner compute, generator compute or total search compute is actually matched.

Timing and independent tests. Separate feedback data, candidate-selection data and final test data. Freeze the chosen generator and learner before final evaluation.

3. Does revising the improver improve later learning?

Comparison. Start from the same learner and improvement procedure. Compare keeping that procedure fixed with revising it, then evaluate their subsequent learning on separate later tasks or cycles.

Keep fixed or disclose. Shared initial agent, update opportunities, data access, task distribution, resource envelope, evaluator and outcome definition. Record any intended difference.

Timing and independent tests. Separate evidence used to construct the revised improver from the later learning stream and its independent tests. Evaluate the subsequent learning effect rather than only the already-selected best artifact.

Complete the plan before execution

Fill templates/experiment_plan.csv or .json. State the target claim, changed object, starting agent, treatment/control, task protocol, update/selection/test data, freeze point, feedback access, matching scope, native metrics, uncertainty procedure, cost boundary and version references. Configuration and metric choices remain with the study author.

Record the actual comparison

Keep the plan, update history, resources and outcomes linked by stable design/run identifiers.

Separate three uses of data

Updating changes the agent or its improvement components. Selection accepts a candidate or checkpoint. Testing estimates performance after those choices. Record when labels and outcomes become available, when memory is written, and the last decision before an untouched test. An online adaptation stream can allow between-case writeback, but should not be called an untouched final test.

Make matching explicit

List the quantities that are actually matched: initial learner/harness, data access, accepted-data volume, learner-training compute, generator/search compute, evaluation and total resources. Preserve mismatches rather than calling all of them "equal compute".

File	What one record represents
experiment_plan	One planned question, treatment/control design, data timeline, scope of matching and intended outcomes.
run_log	One condition and repetition, its versions, input order, memory state, execution status and artifact/result links.
update_log	One update candidate, its parent/new version, available feedback, selector/evaluator and acceptance/rejection decision.
cost_log	One phase or source-summary condition, including failed/rejected candidates and original resource units.
metric_results	One task-native outcome, its value/status, test role, aggregation and reported uncertainty.

Worked cost example from ECG-RSI Table 10

Condition	Candidates	Accepted	Rejected, derived
Full method	16,193	10,000	6,193
Frozen operators	36,723	10,000	26,723
Frozen skills	22,398	10,000	12,398

Accepted volume is equal; candidate-generation counts are unequal. Rejected counts are subtraction, not additional measurements. The example does not fill unavailable token, runtime or monetary costs with zero. Cost and outcome example rows are labelled reported_summary, not actual_run.

Use missing-value statuses explicitly

Distinguish measured/reported values, not_extracted, not_reported_in_examined_source, not_applicable and planned-but-unexecuted records. A missing value is not zero; a value absent from this extraction is not proof it was absent from the source paper. Preserve actual seeds, trials and failures without inventing them.

Worked assessments from reported studies

These are completed reconstructions of published/preprint designs. This project did not execute their experiments.

A. PrivMeSA: persistent memory under a fixed procedure

Treatment. Memory starts empty and retains distilled lessons between encounters.

Control. The same policy processes the same encounter sequence with memory kept empty.

Held fixed. Local policy and remote specialists; distiller, retrieval and memory-gate design.

Sequence and separation. 897 encounters from 200 held-out patients; patient visits remain chronological within a randomized sequence. Writeback follows each encounter without recorded answers.

Scope of matching. Paired encounter sequence and policy. This isolates the memory condition rather than a revised distiller or training procedure.

Native outcomes and cost. Preserve task-native accuracy/privacy measures and specialist-message counts. The case does not establish complete wall-time or monetary matching.

What this comparison supports. Contribution of retained lesson memory under a fixed procedure.

Boundary. Retrospective replay with synthetic identifiers; not prospective care or revised-improver evidence.

Original source, v1; Section 4.1; Appendix E.1 and Table 3. Source locators are retained in worked_designs.json.

B. ECG-RSI: changing synthesis components

Treatment. Full synthesis method with operator and skill updates.

Control. Freeze operator updates or freeze skill updates, then train on the same accepted-question target.

Held fixed. Learner-training protocol, accepted target of 10,000 questions and downstream evaluation; acceptance gates remain fixed.

Sequence and separation. Repair/generate and select training material before learner training; retain separate feedback, selection and external-test roles.

Scope of matching. Accepted volume and learner training are matched. Candidate-generation counts differ, so full search-cost matching is not established.

Native outcomes and cost. Table 10 reports candidate counts 16,193/36,723/22,398 and native S7 scores 43.22/39.29/40.15 for full/frozen-operator/frozen-skill conditions.

What this comparison supports. The changed synthesis components affect data yield and downstream performance in this bounded comparison.

Boundary. A fixed procedure admits generator changes; this does not test a revised improver.

Original source, v1; Section 5.5 and Table 10; Section 3.1. Source locators are retained in worked_designs.json.

Recipe change and the improver boundary

Do not turn a selected-artifact improvement into evidence that the process of making future updates improved.

C. ReVision3D: training-recipe comparison

Treatment. A changed training recipe selects learnable prompts and uses fresh views.

Control. Longer training without the selected recipe, under the reported equal-training-compute comparison.

Held fixed. Common reader start, downstream reading protocol and mask-grounded verifier; the designer and acceptance rule stay fixed.

Sequence and separation. Development data guide revisions; the nominal final test follows frozen choices. Prior use of some nominal test patients is separately disclosed in the source.

Scope of matching. The longer-training control addresses training compute. The comparison does not establish equality of all outer-loop search cost.

Native outcomes and cost. Retain the source-native lesion recall, false positives and J definitions; report recipe-search and training costs separately.

What this comparison supports. A changed training recipe can affect the trained reader under the stated comparison.

Boundary. Retrospective localization with a fixed designer/verifier. Source test-reuse limitations remain visible.

Original source, v2; Section 5.4; Appendix B and Appendix G.2, Table A17. Source locators are retained in worked_designs.json.

D. ReVision3D: a negative revised-improver assessment

Treatment. Observed recipe/perception/inference revisions proposed by a fixed designer.

Control. The reported recipe and component controls do not compare a revised improvement procedure with its original fixed version.

Held fixed. Designer, attribution, verifier, objective and acceptance rule.

Sequence and separation. The source evaluates selected revised artifacts. It does not isolate the playbook or a revised procedure through independent subsequent-learning cycles.

Scope of matching. No matched changed-improver versus frozen-improver comparison is identified in this case.

Native outcomes and cost. No new revised-improver effect or cost value is inferred from the selected-artifact scores.

What this comparison supports. Reported improvements concern revised artifacts under fixed update rules; the later-learning effect of revising those rules is not isolated.

Boundary. The reported comparison does not establish whether a revised improver improves subsequent learning.

Original source, v2; Section 3; Section 5.4; Appendix G.2. Source locators are retained in worked_designs.json.

Use the 85-study matrix and preserve attribution

The matrix joins the approved review's levels, clinical settings, control descriptions, independent-test descriptions and original source locators without reclassification. Its author-check columns are blank. It includes existing contextual comparators and is not a list of 85 demonstrated RSI systems.

Conceptual sources: Liu et al. (2026), component versus improver; Han et al. (2026), level roles; Pu et al. (2026), functional distinction. Credit the primary studies for their methods/results and RSI-HA for this healthcare operationalization and resource.