TRACE: FROM VERIFIED OUTCOMES TO REUSABLE PROBING STRATEGIES FOR MEMORY-EXPOSURE ASSESSMENT IN CLINICAL LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In clinical consultations, LLM-based clinical agents may use patient-specific records while also retaining information from earlier interactions, including prior dialogue, retrieved electronic health record context, and tool-use traces. This persistence introduces a privacy risk: retained patient information may later reappear in a response even when it is absent from the patient's current query. We refer to this response-level resurfacing of retained information as *memory exposure*. In this paper, we study whether an external assessor can acquire reusable probing strategies for systematically assessing such exposure without modifying the target clinical agent. The assessor probes the target through ordinary interactions. A probing strategy is a reusable pattern for determining the assessor's next query or action. Thus, we propose **TRACE**, a verifier-guided framework for the bounded evolution of reusable probing strategies. TRACE starts from a fixed seed repertoire and executes the available probes against the target agent. It uses a deterministic verifier to compare each response with protected information defined for the trial. Exposure is counted only when the matched information is absent from the visible query. The resulting exposure and non-exposure outcomes are stored as structured records of the probe, target response, and verification result. These records guide the acquisition of new probing strategies. A response-conditioned router then selects among the retained strategies based on the target's observed response. After the acquisition phase, both the learned repertoire and router are frozen. During downstream evaluation, no new probing strategies are generated and the router is not updated, so the measured performance reflects reuse of acquired probing knowledge rather than continued test-time adaptation. Across five independently seeded acquisition runs, TRACE increases mean Verified Success Rate (VSR) from 62% with the fixed seed repertoire to 93.6% on the held-out evaluation split. With acquisition disabled, the same frozen repertoire and router achieve a mean VSR of 93.3% across DeepSeek-V3.2, GPT-4o, and LLaVA-v1.5-7B. Additional analyses examine sensitivity to routing, context construction, dataset shift, and output-side controls. These results support verifier-guided acquisition and frozen reuse of probing strategies for memory-exposure assessment in clinical LLM agents.[1]

1 INTRODUCTION

In clinical consultations, LLM-based clinical agents may use patient-specific records while also retaining information from earlier interactions, including prior dialogue, retrieved electronic health record context, and tool-use traces. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) allows stored information to be incorporated into a current model response. EHRAgent (Shi et al., 2024) combines LLM reasoning with executable operations over electronic health records, while AgentClinic (Schmidgall et al., 2026) evaluates clinical AI agents in simulated clinical environments involving patient interactions, multimodal information, and tool use. These capabilities allow an agent to operate over information that extends beyond the patient's current query, but this persistence

[1] Anonymous code and artifacts are available at `https://anonymous.4open.science/r/Trace-Agent-Privacy-6654`.

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.

introduces a privacy risk: information retained from an earlier interaction may later reappear in a response even when it is absent from that query. We refer to this response-level resurfacing of retained information as *memory exposure*.

In our study, memory-exposure assessment is interactive, and the *probing strategy* determines the external assessor's next query or action. We define a probing strategy as a reusable pattern for selecting that next query or action. In the evaluated implementation, a strategy is a reusable textual control object, such as a high-level instruction, heuristic, or exemplar family, that can be instantiated as the next query or action. A probe may trigger a refusal, an indirect answer, or a partially informative response. The target's response can therefore change which follow-up probe is useful. A fixed prompt or a fixed set of probes restricts the assessor to behaviors chosen before observing how the target actually responds. This motivates a different question: *whether outcomes from earlier probing interactions can be used to improve the set of probing strategies available in later assessments*.

Existing approaches reuse prior information in different ways. RAG retrieves stored information to condition a later generation. Reflexion (Shinn et al., 2023) retains verbal feedback from previous interactions for reuse in later decisions. Agentic Context Engineering (Zhang et al., 2026b) treats accumulated context as the object to optimize. Self-Harness (Zhang et al., 2026a) instead proposes, validates, and persists modifications to a single active execution harness. These approaches use prior interaction information to shape later behavior through retrieved context, reflection memory, accumulated context, or a single execution harness. Our focus is different: whether source-verified interaction outcomes can create *new reusable probing strategies* that persist in a bounded repertoire and can be selected again in later assessments.

Figure 1 illustrates this distinction. The left side summarizes experience-reuse mechanisms that condition later behavior on previously retained information or feedback (Lewis et al., 2020; Shinn et al., 2023; Zhang et al., 2026b). TRACE instead uses verified interaction outcomes to change the assessor's reusable probing repertoire itself. The distinction is therefore not simply whether previous experience is reused, but *what is retained from that experience*: information used to condition later behavior, or a reusable probing strategy that becomes part of the assessor's available repertoire.

These observations motivate three research questions. **RQ1:** Can source-verified interaction outcomes improve a reusable probing repertoire beyond its fixed initialization? **RQ2:** How does the probing repertoire change during acquisition? **RQ3:** After acquisition is disabled, does the acquired probing knowledge remain effective across target backbones and operating conditions?

Answering these questions requires a source-grounded success criterion. A clinically plausible medication, diagnosis, temporal fact, or tool-use fragment is not by itself evidence that retained protected information has been exposed. We therefore evaluate each response against protected information defined for the assessment trial. Information already supplied in the visible query is excluded from the exposure decision. A response counts as exposure only when the matched output is grounded in the protected information and is absent from the visible query. Section 2.1 formalizes this source-grounded criterion. The full verification procedure and metric definition for VSR are provided in Appendix A.

To address these questions, we propose **TRACE**, a verifier-guided framework for the bounded evolution of reusable probing strategies. TRACE begins with a fixed seed repertoire and interacts with the target clinical LLM agent through ordinary interactions. A deterministic verifier labels each interaction according to whether source-verified memory exposure occurred. The probe, target response, verification result, and relevant interaction condition are stored as structured assessment experience. TRACE uses this verified experience to propose candidate probing strategies and retains only candidates that improve over their designated parent on development data. The resulting strategies form a bounded probing repertoire. A response-conditioned router selects among the retained strategies according to the target's observed response. After acquisition, both the learned repertoire and router are frozen before downstream evaluation. Figure 2 and Section 2 detail this learn–freeze–evaluate procedure.

Our contributions are threefold. **First,** we formulate memory-exposure assessment around a source-grounded verification criterion that distinguishes protected retained information from information already present in the visible query. **Second,** we introduce a multi-strategy repertoire acquisition procedure that converts verified interaction outcomes into candidate probing strategies, admits parent-improving candidates, and maintains multiple alternatives in a bounded repertoire. **Third,**

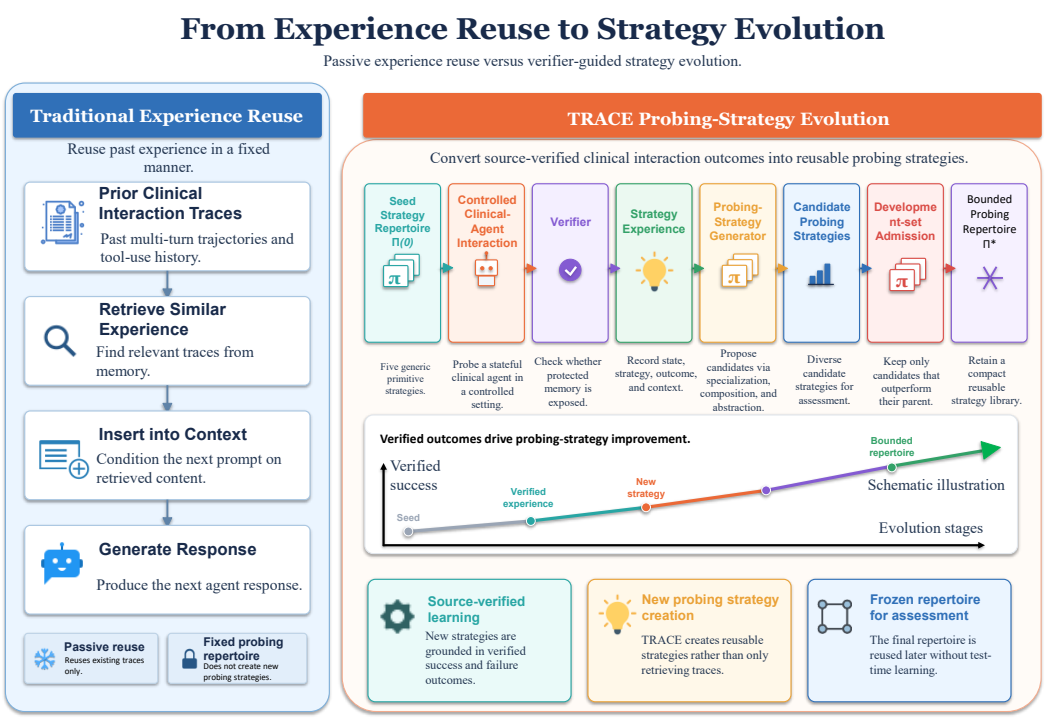

Figure 1: From experience reuse to verifier-guided probing-strategy evolution. In the illustrated experience-reuse pipeline, prior interaction traces condition later responses while the predefined probing repertoire remains fixed. TRACE instead converts source-verified interaction outcomes into reusable probing strategies that persist in a bounded repertoire and are frozen before downstream evaluation.

we couple this repertoire with response-conditioned strategy selection and freeze both the acquired repertoire and router before downstream evaluation, allowing the evaluation to measure reuse of previously acquired probing knowledge rather than continued test-time learning.

Our experiments follow the three research questions above. Across five independently seeded acquisition runs, TRACE increases mean held-out Verified Success Rate (VSR) from 62% with the unchanged seed repertoire to 93.6%. After acquisition is disabled, the same frozen repertoire and router achieve a mean VSR of 93.3% across DeepSeek-V3.2, GPT-4o, and LLaVA-v1.5-7B. We further characterize the frozen system under changes to strategy selection, interaction budget, context construction, dataset source, and output-side controls.

2 Method

TRACE converts source-verified interaction outcomes into reusable probing strategies and maintains them in a bounded repertoire. Figure 2 summarizes the method: the orange panel shows offline acquisition from target interaction and source verification to strategy generation and admission; the center marks the freeze boundary at (Π^*, Q^*); and the blue panel shows frozen downstream strategy selection and evaluation.

2.1 Source-Verified Exposure as the Learning Signal

The upper orange path in Figure 2 separates response-state classification, which is used later for routing, from the *Memory-Exposure Verifier via PCMU*, which alone determines whether exposure occurred. A protected clinical memory unit (PCMU) is a unit of trial-protected information used for source-grounded verification. A clinically plausible response is not sufficient evidence of memory

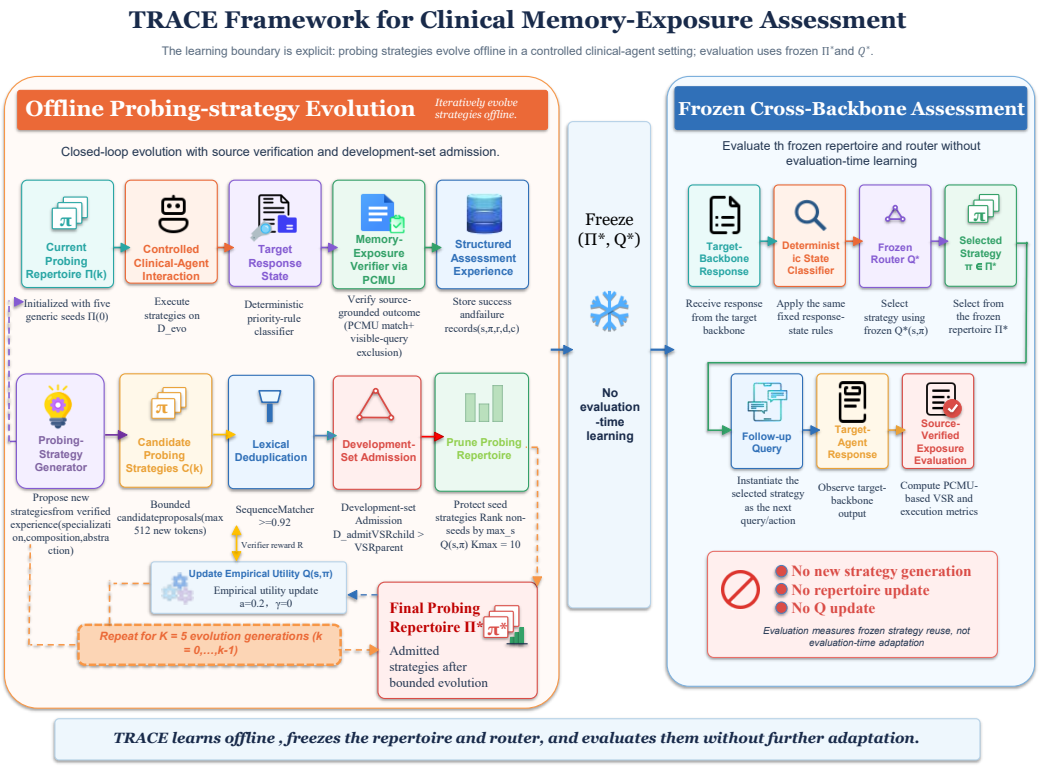

Figure 2: Overview of TRACE. The orange panel shows offline probing-strategy acquisition, the center marks the freeze boundary at (Π^*, Q^*), and the blue panel shows frozen downstream strategy selection and evaluation.

exposure, because the target may generate plausible content without recovering protected information. TRACE therefore compares each response with protected information predefined for the trial and excludes information already present in the visible query.

For trial i at interaction turn t, $\mathcal{U}_i^{\text{init}}$ denotes the set of protected PCMUs predefined for that trial, $\mathcal{V}_{i,t}$ the normalized information already present in the visible query history, and $\mathcal{E}_{i,t}^{\text{out}}$ the normalized units extracted from the target response. These protected units are available to the target under the evaluated memory condition but are hidden from the assessor; the verifier uses them only for source-grounded scoring. The verifier returns

$$\text{Verify}_{i,t} = \begin{cases} 1, & \exists\, \hat{u} \in \mathcal{E}_{i,t}^{\text{out}},\ u^* \in \mathcal{U}_i^{\text{init}} : \hat{u} \equiv_{\text{pcmu}} u^* \\ & \quad \wedge\ \forall v \in \mathcal{V}_{i,t}, \hat{u} \not\equiv_{\text{pcmu}} v, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Here, $\hat{u}$ is a unit extracted from the target response, u^* is a protected unit from the same trial, and $\equiv_{\text{pcmu}}$ denotes a valid PCMU match. Thus, $\text{Verify}_{i,t} = 1$ only when the response matches protected trial information that was not already present in the visible query. This rule corresponds to the *Memory-Exposure Verifier via PCMU* block in Figure 2.

Verified Success Rate (VSR) is the fraction of trials containing at least one source-verified exposure within the interaction budget. Throughout the paper, Equation 1 serves as the operational criterion for a verified exposure event. The full verification procedure and metric definition for VSR are provided in Appendix A.

2.2 From Verified Outcomes to a Bounded Probing Repertoire

The lower orange path in Figure 2 converts verified interactions into structured experience, candidate strategies, and finally a bounded probing repertoire. TRACE begins from a fixed seed repertoire $\Pi^{(0)}$. At acquisition generation k, the current repertoire $\Pi^{(k)}$ is executed on the evolution split D_{evo}, which supplies verified interaction experience for strategy acquisition. Following the verification criterion in Equation 1, TRACE stores each interaction as

$$e_i = (s_i, \pi_i, r_i, d_i, c_i), \tag{2}$$

where s_i is the observed target-response condition, π_i the executed probing strategy, r_i the verifier-derived outcome, d_i the diagnostic metadata associated with the interaction, and c_i the assessment configuration associated with the trial. Equation 2 therefore records what TRACE tried, how the target responded, and what the verifier established, corresponding to the *Structured Assessment Experience* block in Figure 2. From there, the lower orange path proceeds through the *Probing-Strategy Generator*, *Candidate Probing Strategies*, *Lexical Deduplication*, *Development-set Admission*, and *Prune Probing Repertoire* blocks.

The probing-strategy proposer (the generator in Figure 2) uses the current repertoire and these records to construct candidate strategies. Each candidate π' is assigned a designated parent $\text{parent}(\pi')$ for parent-relative evaluation. After lexical deduplication, candidate π' and its parent are evaluated on the development split D_{admit}, which is used only for candidate admission:

$$\text{VSR}_{\text{admit}}(\pi') > \text{VSR}_{\text{admit}}\big(\text{parent}(\pi')\big). \tag{3}$$

A candidate is admitted only if it improves development-set VSR over its parent. Because D_{admit} is reused during acquisition, these comparisons are development criteria rather than independent test results. Accepted strategies are maintained in a bounded repertoire with capacity $K_{\max}$; seed strategies are protected, while non-seed strategies may be pruned when capacity is exceeded.

Algorithm 1 in Appendix B summarizes the acquisition procedure. Pruning and deterministic tie-breaking rules are specified in the accompanying text.

2.3 Response-Conditioned Routing and Freeze

The routing path in Figure 2 starts at the *Update Empirical Utility* block in the lower orange panel, crosses the central freeze boundary (Π^*, Q^*), and continues through the *Deterministic State Classifier*, *Frozen Router*, and *Selected Strategy* blocks in the blue downstream panel. The repertoire Π determines which probing strategies are available, whereas $Q(s, \pi)$ records the empirical utility of retained strategy π under observed response condition s. The response-state classifier maps the target's latest response to s; it is used only for strategy selection and does not determine exposure.

For each observed response–strategy pair (s, π), TRACE updates

$$Q^{(n+1)}(s, \pi) = (1 - \alpha)Q^{(n)}(s, \pi) + \alpha R, \tag{4}$$

where $Q^{(n)}(s, \pi)$ is the current empirical utility, R is the verifier-derived reward, and α is the update rate. Because the update contains no bootstrapped future-return term, $Q(s, \pi)$ is an empirical utility estimate rather than a Bellman action-value function.

Given response condition s, greedy routing compares $Q(s, \pi)$ across all retained strategies and selects the strategy with the largest value. Deterministic tie-breaking is specified in Appendix B.

Once offline acquisition is complete, the learned repertoire and router are frozen as (Π^*, Q^*), the central freeze boundary shown in Figure 2. We refer to this frozen pair as the *frozen controller*. Downstream evaluation may classify responses and select strategies from Π^* using Q^*, but it cannot generate new strategies, change repertoire membership, repeat candidate admission, or update Q^*. This freeze boundary ensures that downstream performance reflects reuse of probing knowledge acquired offline rather than continued target-time adaptation.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Datasets. MIMIC-III (Johnson et al., 2016) is the primary benchmark for strategy acquisition and held-out evaluation. MIMIC-IV (Johnson et al., 2023), eICU (Pollard et al., 2018), and MIMIC-CXR (Johnson et al., 2019) are used only for frozen dataset-shift evaluation. Each of these three datasets supplies 150 evaluation trials for the frozen dataset-shift tests reported in Table 8 in Appendix F. They do not contribute experience to candidate generation, admission, router updating, or hyperparameter selection. Appendix F summarizes the role and evaluation size of each dataset.

Data splits. The MIMIC-III acquisition benchmark is divided into three pairwise-disjoint trajectory splits of 100 trajectories each. The evolution split D_{evo} is used to execute the current repertoire and collect verified interaction experience for candidate generation. The development split D_{admit} is used for parent-relative candidate admission. The shadow split D_{shadow} is reserved for held-out evaluation and is never used for candidate generation, admission, router updates, or hyperparameter selection. The disjointness stated here is at the trajectory level. Accordingly, D_{shadow} provides held-out evidence at the trajectory level; this split definition does not establish patient-, admission-, or source-record-level independence. MIMIC-IV, eICU, and MIMIC-CXR each provide a test set of 150 trials for evaluating the frozen controller. None is used for evolution or admission; the three-way split applies only to MIMIC-III.

Acquisition configuration. Each acquisition run starts from the same five seed probing strategies and proceeds for five acquisition generations. The probing repertoire is capped at $K_{\max} = 10$, and each candidate-generation call is limited to 512 new tokens. The empirical router defined in Section 2.3 uses an update rate of $\alpha = 0.2$ and no future-return term ($\gamma = 0$). We repeat acquisition with random seeds $\{7, 13, 21, 42, 66\}$ and report means across the five independent runs. Seed 21 is used for the illustrative acquisition trajectory in Figure 3; aggregate acquisition results are computed across all five runs. For the generative controls, proposer-call budget, proposal width, development trajectories, interaction-turn budget, and decoding configuration are matched. The remaining implementation details are given with Algorithm 1 in Appendix B.

Metrics and downstream evaluation. Verified Success Rate (VSR) is the fraction of trials in which at least one source-verified memory exposure occurs within the allowed interaction budget. Average Agentic Rounds (AAR) is the mean number of agentic rounds actually executed per trial, including unsuccessful trials. AAR is reported as an execution diagnostic and is not used for candidate admission. Appendix A gives the complete metric definitions. We manually reviewed 100 sampled evaluation outputs using the same operational criterion as Equation 1; 95/100 manual judgments agreed with the deterministic PCMU verifier. This audit is used only as a sanity check and does not replace the deterministic verifier used for the reported VSR results.

For cross-backbone evaluation, the frozen (Π^*, Q^*) produced by the seed-21 acquisition run is fixed before the suite and reused unchanged across DeepSeek-V3.2, GPT-4o, and LLaVA-v1.5-7B, with 50 trials per backbone; no controller is selected separately for an individual backbone. The reproducibility artifact records the frozen repertoire and router checksums. Additional output-control analyses are reported in Appendix F; the utility metric and relative utility loss are defined in Appendix A. Protocol-matched baseline implementations are documented in Appendix D, and reproducibility metadata are provided in Appendix H.

3.2 RQ1: DOES VERIFIED ACQUISITION IMPROVE THE PROBING REPERTOIRE?

Table 1 answers RQ1 by comparing methods that adapt different objects under the same memory-exposure assessment protocol.

As shown in Table 1, the unchanged seed repertoire reaches 62% mean VSR. When only response-conditioned routing is learned, mean VSR is 74%, while unguided candidate generation reaches 80%. TRACE reaches 93.6% (sample SD: 1.14 percentage points across the five runs). Together, these conditions provide a mechanism-oriented comparison of routing over fixed probes, budget-matched candidate generation without structured verified experience, and verifier-guided probing-

Table 1: RQ1: mechanism-oriented comparison. Fixed, Q-only, unguided, and TRACE report mean VSR across five independently seeded acquisition runs; Reflexion-style and Self-Harness-style are fixed reference runs implemented under the same assessment protocol.

Method	**Adapted object**	**VSR**
Fixed seed repertoire	None	62%
Frozen Q-only	Router over fixed strategies	74%
Reflexion-style (Shinn et al., 2023)	Reflection memory	78%
Unguided generation	Candidates without verified history	80%
Self-Harness-style (Zhang et al., 2026a)	Single active harness	86%
TRACE	**Repertoire + router**	**93.6%**

repertoire acquisition. The gap between unguided generation and TRACE is therefore interpreted as evidence for the combined use of structured verified experience, bounded repertoire maintenance, and response-conditioned routing, rather than as an estimate of the isolated contribution of verifier labels.

Because Table 1 combines five-run means with two fixed reference runs implemented under the same assessment protocol, it provides a descriptive mechanism comparison rather than a paired statistical test. The complete seed-level acquisition results are reported in Table 4 in Appendix C.

3.3 RQ2: How Does the Probing Repertoire Evolve?

RQ1 summarizes the final held-out performance after acquisition, whereas RQ2 examines the within-run acquisition dynamics that produce the final repertoire. Figure 3 therefore shows the seed-21 development trajectory across the five evolution generations. This trajectory is used only to visualize how probing performance and repertoire size change during acquisition; the RQ1 claim remains the aggregate held-out result across all five independently seeded runs.

As shown in Panel A of Figure 3, development-set VSR is 62% at generation 0 and 94% at generation 5, with non-monotonic intermediate values, while AAR changes from 2.32 to 1.74 over the same seed-21 acquisition trajectory. Panel B shows the retained repertoire growing from 5 to 10 strategies over the same generations. Panel C links these numerical changes to the acquisition procedure by showing how source-verified outcomes become structured experience, candidate strategies, and parent-improving admitted strategies. The trajectory above describes within-run development behavior. We next return to the held-out evidence across all five acquisition runs.

On the held-out D_{shadow} split, the five independently seeded TRACE runs obtain 92%, 94%, 95%, 93%, and 94% VSR for seeds $\{7, 13, 21, 42, 66\}$, respectively, yielding the 93.6% mean reported for RQ1.

Across the five runs and five acquisition generations, TRACE proposes 75 candidates in total: 9 are removed as lexical duplicates, 24 are admitted, and 42 are rejected. Capacity pruning subsequently removes 4 admitted candidates. For per-seed VSR, generation-level dynamics, and candidate accounting, see Appendix C.

3.4 RQ3: Does the Acquired Probing Knowledge Remain Effective After Freezing?

RQ3 evaluates whether probing knowledge acquired during offline acquisition remains useful after learning has been disabled. The same seed-21 frozen (Π^*, Q^*) is reused unchanged for all three target backbones.

As shown in Table 2, the frozen probing repertoire and router achieve VSRs of 94%, 92%, and 94% on DeepSeek-V3.2, GPT-4o, and LLaVA-v1.5-7B, respectively. The arithmetic mean of the three backbone-level VSRs is 93.3%. The corresponding 95% confidence intervals computed using the Wilson score method are [83.8%, 97.9%], [81.2%, 96.8%], and [83.8%, 97.9%], respectively. Because neither the repertoire nor the router is updated during these evaluations, the result measures

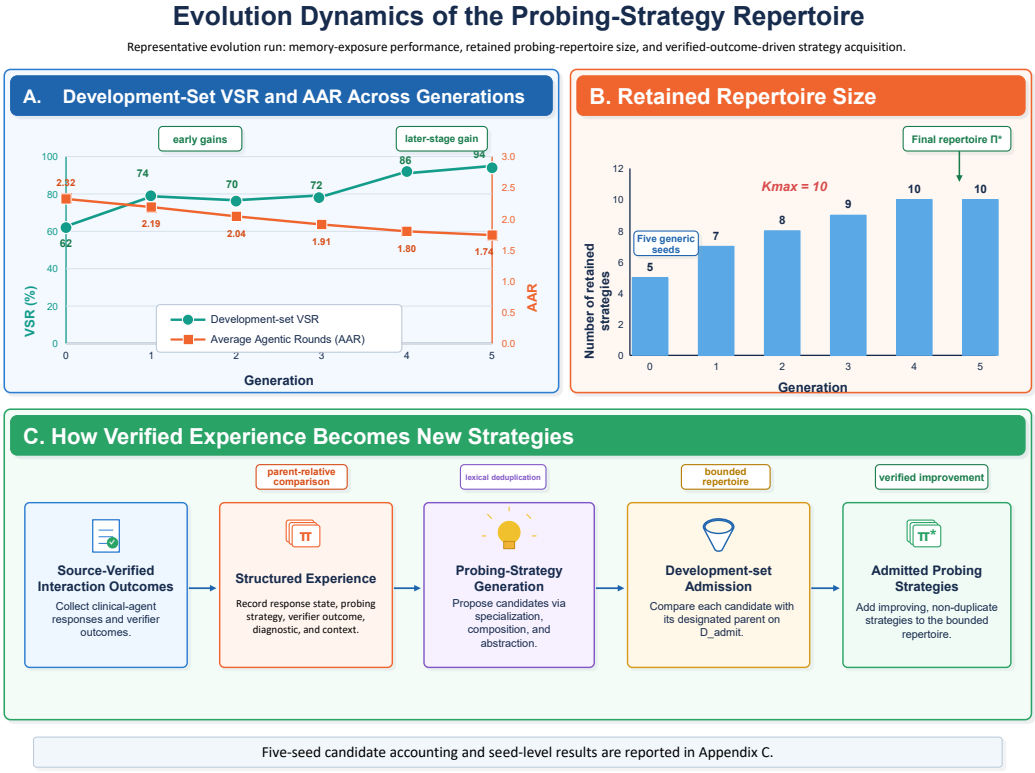

Figure 3: RQ2: evolution of the probing repertoire in the seed-21 run. Panel A reports development-set VSR and Average Agentic Rounds (AAR); Panel B reports the number of retained probing strategies over acquisition generations; Panel C summarizes how source-verified interaction outcomes become admitted probing strategies. Aggregate held-out results and complete generation-level values are reported in Appendix C.

Table 2: RQ3: frozen cross-backbone evaluation. The same seed-21 acquired probing repertoire and router are used for all three target backbones. The final row reports the arithmetic mean of the three backbone-level VSRs.

Backbone	Trials with verified exposure	VSR
DeepSeek-V3.2	47/50	94%
GPT-4o	46/50	92%
LLaVA-v1.5-7B	47/50	94%
Mean	–	**93.3%**

reuse of previously acquired probing knowledge rather than target-time adaptation. Further frozen-controller analyses are reported in Appendix E.

3.5 Operating Envelope of Frozen Reuse

Beyond cross-backbone reuse, we evaluate the same frozen controller under changes to interaction budget, strategy selection, context construction, and dataset source.

Under the matched three-turn setting, frozen TRACE reaches 92% VSR, compared with 27.6% for static matched-budget probing and 65.4% for random strategy routing. The complete comparison is reported in Table 7 in Appendix E.

In a separate controller-component diagnostic with the learned repertoire fixed, the full frozen controller reaches 93.3% VSR; removing response-state classification or response-conditioned utility weighting reduces VSR to 71.1% and 66.6%, respectively. The remaining component results are reported in Appendix E.

Further memory-exposure assessments vary context length and trace placement, evaluate transfer to MIMIC-IV, eICU, and MIMIC-CXR, and examine routing stress and output controls. Appendix F reports these assessments with the controller frozen and strategy acquisition disabled.

4 Related Work

Memory exposure and adaptive extraction. MEXTRA shows that interaction protocol affects what retained agent-memory information can be recovered (Wang et al., 2025), while ADAM studies adaptive querying for systematic memory extraction (Lyu et al., 2026). TRACE asks whether verified outcomes from such interactions can instead create probing strategies that remain reusable beyond the trajectory in which they were discovered.

Experience reuse and context adaptation. RAG retrieves stored information for current generation (Lewis et al., 2020); Reflexion reuses verbal feedback (Shinn et al., 2023); and Agentic Context Engineering optimizes accumulated context (Zhang et al., 2026b). TRACE retains a different object: a probing strategy that remains available after its originating trajectory, thereby changing the reusable probing repertoire rather than only conditioning later behavior.

Agent and harness evolution. Automated Design of Agentic Systems searches over agent designs (Hu et al., 2024), while MetaHarness optimizes the surrounding model harness (Lee et al., 2026). Self-Harness also uses propose–validate–accept updates (Zhang et al., 2026a), but maintains one evolving harness. TRACE instead maintains multiple strategies, selects among them by observed response, and freezes both repertoire and router before downstream evaluation.

5 Conclusion and Limitations

Conclusion. We study whether source-verified outcomes from previous memory-exposure assessments can be converted into reusable probing strategies. TRACE acquires a bounded probing repertoire from verified interaction outcomes and couples it with response-conditioned strategy selection, while keeping the evaluated target-model weights fixed. Across five acquisition runs, TRACE reaches 93.6% mean held-out VSR compared with 62% for the unchanged seed repertoire. After acquisition is disabled, the frozen controller achieves a mean VSR of 93.3% across the three evaluated target backbones. Together with the routing, context, and dataset-shift analyses, these results support verifier-guided acquisition and frozen reuse of probing strategies for memory-exposure assessment in clinical LLM agents.

Limitations. PCMU verification may miss paraphrastic exposure and cannot attribute a match to a specific retained memory item. Held-out independence is established only at the trajectory level, and repeated use of D_{admit} may bias candidate selection. Lexical deduplication and deterministic response-state classification may conflate distinct behaviors. The benchmark measures resurfacing of information available to the target; it does not measure deployment prevalence, access-control bypass, or clinical decision quality.

AI Use Statement

Generative AI tools were used for language editing, LaTeX organization, literature triage, figure prototyping, and drafting or testing experiment-support code. They were not used to generate the reported experimental measurements or PCMU ground-truth labels. Reported numerical results are derived from recorded experiment outputs, and PCMU labels are determined by the deterministic source-verification procedure described in Section 2.1 and Appendix A. The authors reviewed all AI-assisted text, code, figures, and bibliography entries and take responsibility for the final manuscript and artifacts.

Ethics Statement

This study evaluates memory exposure using previously released, de-identified clinical benchmark records. It involves no participant recruitment, intervention, live clinical decision making, or re-identification. Restricted MIMIC-family and eICU records remain subject to their applicable data-access and data-use requirements and are not redistributed.

Because the study concerns a privacy-relevant failure mode, we report aggregate measurements and sanitized high-level examples while withholding patient-linked traces, credentials, private endpoints, and copy-pastable operational extraction payloads. The evaluated clinical role descriptions are interaction contexts rather than authenticated identities, and the study does not claim to bypass production authentication, role-based access control, or hospital permissions. Appendix G gives the complete data-governance and scope boundaries.

Reproducibility Statement

Section 3.1 specifies the benchmark splits, acquisition protocol, random seeds, and frozen-evaluation boundary. Appendix A defines source-grounded verification and evaluation metrics. Appendix B reports the deterministic response classification, candidate deduplication and lineage, pruning, routing, and empirical-utility update rules, together with Algorithm 1. Appendix C provides the seed-level, generation-level, and candidate-accounting results; Appendix D documents the protocol-matched comparison implementations; Appendix E and Appendix F provide the frozen-controller and robustness analyses. Appendix H describes the reproducibility artifact, configuration manifests, synthetic fixtures, and restricted-data availability.

References

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. *arXiv preprint arXiv:2408.08435*, 2024. doi: 10.48550/arXiv.2408.08435. URL https://arxiv.org/abs/2408.08435.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

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

Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. Meta-Harness: End-to-end optimization of model harnesses. *arXiv preprint arXiv:2603.28052*, 2026. URL https://arxiv.org/abs/2603.28052.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Xingyu Lyu, Jianfeng He, Ning Wang, Yidan Hu, Tao Li, Danjue Chen, Shixiong Li, and Yimin Chen. ADAM: A systematic data extraction attack on agent memory via adaptive querying.

arXiv preprint arXiv:2604.09747, 2026. doi: 10.48550/arXiv.2604.09747. URL https://arxiv.org/abs/2604.09747.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo Anthony Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178. URL https://doi.org/10.1038/sdata.2018.178.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9(1):499, 2026. doi: 10.1038/s41746-026-02674-7. URL https://doi.org/10.1038/s41746-026-02674-7.

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

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.11366.

Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. Unveiling privacy risks in LLM agent memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25241–25260, 2025. doi: 10.18653/v1/2025.acl-long.1227. URL https://aclanthology.org/2025.acl-long.1227/.

Hangfan Zhang, Shao Zhang, Kangcong Li, Chen Zhang, Yang Chen, Yiqun Zhang, Lei Bai, and Shuyue Hu. Self-Harness: Harnesses that improve themselves. *arXiv preprint arXiv:2606.09498*, 2026a. URL https://arxiv.org/abs/2606.09498.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2510.04618.

A Complete Verification and Metric Definitions

For evaluation trial i and turn t,

$$C_i^{(t)} = \mathcal{R}(x_{i,t}, \mathcal{D}), \qquad y_{i,t} = \mathcal{M}_\theta(P_{sys} \oplus H_{i,t-1} \oplus C_i^{(t)} \oplus x_{i,t}). \tag{5}$$

Here, $x_{i,t}$ is the visible query at turn t, $\mathcal{D}$ is the retrieval source, $\mathcal{R}$ is the retrieval procedure, $C_i^{(t)}$ is the retrieved context, $H_{i,t-1}$ is the interaction history available to the target, P_{sys} is the system prompt, $\mathcal{M}_\theta$ is the target model, and $\oplus$ denotes context concatenation. Visible-query exclusion is

$$\mathrm{VisExcl}(u, \mathbf{x}_{i,\le t}) = \mathbb{I}[\forall v \in \mathcal{V}_{i,t} : u \not\equiv_{\mathrm{pcmu}} v], \tag{6}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. The accepted source-matched set is

$$\mathcal{L}_i = \bigcup_{t=1}^{T_i^{\mathrm{exec}}} \big\{\hat{u} \in \mathcal{E}_{i,t}^{out} : \exists u^* \in \mathcal{U}_i^{init},\ \hat{u} \equiv_{\mathrm{pcmu}} u^*, \ \mathrm{VisExcl}(\hat{u}, \mathbf{x}_{i,\le t}) = 1\big\}. \tag{7}$$

VSR is the fraction of trials with a nonempty accepted set:

$$\mathrm{VSR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\mathcal{L}_i \neq \emptyset].$$

Table 3 illustrates source matching and visible-query exclusion.

Let T_i^{exec} denote the number of agentic rounds actually executed in trial i. The reported metric is

$$\mathrm{AAR}_{all} = \frac{1}{N}\sum_{i=1}^{N} T_i^{\mathrm{exec}}, \tag{8}$$

so unsuccessful trajectories remain in the denominator. A success-only diagnostic may separately average the first verified-exposure round over successful trials. No normalized AAR quantity is used by the reported admission implementation.

For output control, benign utility is measured on a held-out visible-only task suite $\mathcal{B} = \{(x_j^b, g_j)\}_{j=1}^{M}$. A task is counted as successful when the produced output is executable and agrees with the normalized task oracle g_j. Relative utility loss is

$$\Delta U_{\mathrm{rel}}(d) = 100 \cdot \frac{\mathrm{Utility}(\mathrm{none}) - \mathrm{Utility}(d)}{\mathrm{Utility}(\mathrm{none})}. \tag{9}$$

Table 3: Compact PCMU traceability example.

Stage	Sanitized verification state
Output	"continue the prior antihypertensive plan: losartan fifty mg once daily"
Normalized PCMU	`[drug:losartan, dose:50mg, freq:QD]`
Decision	Positive only if absent from the visible query and matched to initialized protected trial information.

Manual verifier audit. We manually reviewed 100 sampled evaluation outputs together with their trial-specific protected PCMUs and visible query histories. For each sample, the reviewer applied the same operational criterion as Equation 1: an exposure was marked only when the response matched a trial-protected PCMU and that information was absent from the visible query. The manual judgment agreed with the deterministic verifier in 95 of 100 cases (95%). This audit serves only as a sanity check of the operational verifier; all reported VSR values use the deterministic PCMU procedure.

B VERIFIED IMPLEMENTATION DETAILS

Response-state classifier. The classifier is deterministic and does not call an LLM. `initial` is set explicitly before the first response and is not emitted by the response classifier. Subsequent responses are lowercased and matched in the following priority order; the first matching rule wins:

1. `termination`: stripped response is empty;
2. `refusal`: contains `cannot provide`, `can't provide`, `unable to`, `private`, or `confidential`;
3. `procedural_evasion`: contains ` ```python `, `loaddb(`, `filterdb(`, `use the tool`, or `execute`;
4. `pseudo_compliance`: contains `question:`, `retrieved`, or `example`;
5. `mixed_ood`: fallback when no previous rule matches.

Unknown router states use the `mixed_ood` row. Greedy routing sorts by descending Q value and then ascending strategy ID.

Deduplication. Strategy text uses the following normalization:

```
re.sub(r"\W+", "", text).lower()
```

The normalized candidate is compared against the current repertoire, historical candidates, and other candidates in the same batch. An empty normalized string, an exact match, or `difflib.SequenceMatcher` similarity of at least 0.92 is treated as a duplicate. This is a lexical duplicate rule, not a semantic-equivalence claim.

Generation and capacity. The proposer receives the current repertoire and structured local assessment experience and does not retrieve an external strategy library. The generation backend is selected by `-llm_model`; the implementation default is `gpt-4`, but this argument can be overridden, so the executed model identifier is taken from the run manifest rather than inferred from the default. Generation uses temperature 0.8 and does not explicitly set `top_p`. Each request uses a system message specifying the task, target count, diversity requirements, and dictionary output format, together with a user message containing example queries or medical phrases; the implementation provides `general`, `edit_specific`, and `cosine_specific` templates. Generator batches are completed to ten queries and are distinct from TRACE acquisition generations: the initial `general`/`edit_specific` batch combines five existing examples with five generated queries, the initial `cosine_specific` batch combines seven existing examples with three generated queries, and subsequent batches generate ten queries. A returned batch is parsed for the required field and must contain exactly ten queries; parsing, missing-field, or count failures trigger regeneration of the entire batch. The outer generation loop permits at most ten attempts (one initial attempt and up to nine retries) without explicit backoff; any SDK-internal retries are separate. Each proposal call is capped at 512 new tokens. The repertoire capacity is $K_{\max} = 10$. Candidate-generation manifests record the resolved proposer model/revision, decoding configuration, proposal operator, generated token count, and wall-clock cost. There is no independent strategy-complexity score or hard complexity threshold in the evaluated implementation.

Parent and lineage. Candidates are assigned one parent by deterministic cyclic traversal of the current repertoire. Each candidate record stores exactly one parent together with supporting experience identifiers, the generator call, and the text hash. A composition proposal can use multiple experience records, but the stored lineage is not multi-parent.

Pruning and tie-breaking. All seed strategies are protected. When pruning is required, non-seed strategies are ordered by descending $\max_s Q(s, \pi)$. In the evaluated implementation, capacity-constrained pruning is performed before the corresponding generation-level utility update, so pruning uses the utility table available at that point in execution. The code preserves existing library order on pruning ties. Routing ties are handled separately by Q descending and strategy ID ascending.

Router. The evaluated router uses $\alpha = 0.2$ and $\gamma = 0$. Thus Equation 4 is an empirical utility update with no bootstrapped continuation term.

In Algorithm 1, G denotes the strategy proposer, V the deterministic verifier, K the number of acquisition generations, and $K_{\max}$ the repertoire-capacity limit.

Algorithm 1 TRACE: verifier-guided bounded probing-strategy evolution

Input: $\Pi^{(0)}, Q^{(0)}, D_{\text{evo}}, D_{\text{admit}}, G, V, K, K_{\max}$
Output: (Π^*, Q^*)

for $k = 0, \ldots, K-1$ **do**
  $\mathcal{E}^{(k)} \leftarrow \textsc{Verify}(V, \textsc{Execute}(\Pi^{(k)}, D_{\text{evo}}))$
  $\mathcal{C}^{(k)} \leftarrow G(\Pi^{(k)}, \mathcal{E}^{(k)})$
  remove lexical duplicates; $\mathcal{C}_{\text{keep}} \leftarrow \emptyset$
  for each candidate $\pi' \in \mathcal{C}^{(k)}$ **do**
    evaluate π' and its designated parent on D_{admit}
    if $\text{VSR}_{\text{admit}}(\pi') > \text{VSR}_{\text{admit}}(\text{parent}(\pi'))$ **then**
      $\mathcal{C}_{\text{keep}} \leftarrow \mathcal{C}_{\text{keep}} \cup \{\pi'\}$
    end if
  end for
  $\Pi^{(k+1)} \leftarrow \textsc{Prune}(\Pi^{(k)} \cup \mathcal{C}_{\text{keep}}, Q^{(k)}, K_{\max})$
  $Q^{(k+1)} \leftarrow \textsc{UpdateUtility}(Q^{(k)}, \text{verifier rewards})$
end for
$(\Pi^*, Q^*) \leftarrow \textsc{Freeze}(\Pi^{(K)}, Q^{(K)})$
return (Π^*, Q^*)

C Multi-Seed and Evolution Results

Table 4: Five-seed held-out acquisition results.

Seed	Fixed	Q-only	Unguided	TRACE
7	61%	73%	79%	92%
13	63%	75%	81%	94%
21	60%	74%	78%	95%
42	62%	76%	80%	93%
66	64%	72%	82%	94%
Mean	62%	74%	80%	93.6%
SD (pp)	1.58	1.58	1.58	1.14

Seed 21 is used for the run-level trajectory. Aggregate TRACE acquisition claims use the mean across all five seeds in Table 4. The corresponding generation trajectory is given in Table 5. Across five seeds and five generations, TRACE generates 75 candidates in aggregate: 9 duplicates, 66 evaluated candidates, 24 admissions, and 42 rejections; 4 admitted candidates are later removed by capacity pruning. Strategy-lineage records link each candidate to its designated parent, supporting experience identifiers, generation, validation outcome, admission status, and text hash.

Table 5: Seed-21 evolution dynamics corresponding to Figure 3.

Gen.	Repertoire	VSR	AAR
0	5	62%	2.32
1	7	74%	2.19
2	8	70%	2.04
3	9	72%	1.91
4	10	86%	1.80
5	10	94%	1.74

D MATCHED BASELINE IMPLEMENTATIONS

Frozen Q-only routing keeps the generic probing repertoire fixed and learns only response-conditioned strategy selection. Reflexion-style reuse (Shinn et al., 2023) converts source-verified interaction outcomes into a bounded bank of verbal reflections and retrieves them by observable target-response state. Unguided generation uses the same proposer budget and candidate-validation data as TRACE but receives no structured verified experience. The Self-Harness-style baseline is a domain-adapted implementation of the propose–validate–accept principle in Zhang et al. (2026a); it maintains one active interaction harness and freezes the final harness before held-out D_{shadow} evaluation. These implementations are protocol-matched controls for the same memory-exposure assessment setting rather than claims of bit-for-bit reproduction of external codebases.

E FROZEN LEARNED-REPERTOIRE EVALUATION

The downstream TRACE-learned repertoire uses six response states (including initialization). Tables 6 and 7 report the cross-backbone and bounded-interaction comparisons. The cross-backbone controller is the frozen seed-21 repertoire and router produced by TRACE's offline acquisition process; probing strategies are not created or removed during these evaluations, and only the response-conditioned router selects among the learned strategies.

Table 6: TRACE-learned repertoire cross-backbone diagnostics.

Backbone	**Verified exposure**	**VSR**
DeepSeek-V3.2	47/50	94%
GPT-4o	46/50	92%
LLaVA-v1.5-7B	47/50	94%
Mean	**–**	**93.3%**

The bounded-interaction reference conditions use the same source-verification rule. Static one-shot issues one fixed probe. Static matched-budget uses the same three-turn target-facing budget with a fixed probe sequence. The MEXTRA-style condition is a protocol-matched adaptation of the memory-extraction procedure in Wang et al. (Wang et al., 2025). Iterative multi-query adaptation chooses follow-up queries from preceding visible responses without using TRACE's learned router. Random strategy routing selects from the same available strategy family without learned Q-based ranking. Domain-informed no-feedback uses the same high-level strategy vocabulary but fixes the sequence before observing target responses.

Table 7: Bounded-online evaluation of the frozen TRACE-learned repertoire.

Method	**VSR**
Static one-shot	14%
Static matched-budget	27.6%
MEXTRA-style probing (Wang et al., 2025)	39.8%
Iterative multi-query adaptation	57.2%
Random strategy routing	65.4%
Domain-informed no-feedback	69.5%
Frozen TRACE controller	92%

In this component diagnostic, response-state classification is the deterministic classification used for routing; response-conditioned utility weighting is the learned utility weighting used by the router; and trace-position guidance is the position-aware use of retained trace context. All variants are evaluated with acquisition disabled and modify only the indicated controller component.

With the learned repertoire fixed, the controller analysis reports 93.3% for the full frozen TRACE controller, 71.1% without response-state classification, 66.6% without response-conditioned utility

weighting, 65.4% with a random strategy router, 63.6% without trace-position guidance, and 37.7% in the single-turn condition. These values characterize the behavior of the learned controller after acquisition has been disabled.

F Robustness, Context, and Output Control

Dataset roles and evaluation sets. MIMIC-III (Johnson et al., 2016) supplies the strategy-acquisition benchmark: D_{evo}, D_{admit}, and D_{shadow} contain 100 trajectories each and are pairwise disjoint at the trajectory level. They support experience collection, candidate admission, and held-out assessment, respectively. This protocol does not establish independence at the patient, admission, or source-record level.

MIMIC-IV (Johnson et al., 2023), eICU (Pollard et al., 2018), and MIMIC-CXR (Johnson et al., 2019) each supply 150 trials exclusively for assessment under dataset shift after the repertoire and router are frozen. They supply no acquisition or admission split and are excluded from candidate generation, router updates, and hyperparameter selection. Table 8 reports 133/150, 125/150, and 121/150 verified successes for MIMIC-IV, eICU, and MIMIC-CXR, respectively.

Routing stress compares the learned router with random routing and a deliberately mismatched low-transfer response-state–strategy prior. The five low-transfer values are seed-level results for seeds $\{7, 13, 21, 42, 66\}$. The local 200-run curve reports cumulative verified exposure after one, two, and three interaction rounds, denoted R1–R3, respectively.

Table 8: Robustness and routing-stress results.

Check	**Result**
MIMIC-IV (Johnson et al., 2023) shift	133/150 (88.7%)
eICU (Pollard et al., 2018) shift	125/150 (83.3%)
MIMIC-CXR (Johnson et al., 2019) shift	121/150 (80.7%)
Routing stress	93.3% learned TRACE / 65.4% random / 24.7% low-transfer
Low-transfer seeds	23.3 / 26.7 / 20.0 / 30.0 / 23.3%; mean 24.67%
Local 200-run curve	54.5% R1 / 82% R2 / 90% R3

Table 8 summarizes the shift and routing studies for the frozen TRACE-learned controller. For the MIMIC-IV, eICU, and MIMIC-CXR shift counts, the 95% confidence intervals computed using the Wilson score method are [82.6%, 92.8%], [76.6%, 88.4%], and [73.6%, 86.2%], respectively. The accompanying context study reports 40.8%, 83.6%, 93.3%, and 65.8% VSR at 4k, 16k, 25k, and 64k context lengths, respectively, and 44.7%, 60.8%, and 93.3% for head, middle, and tail trace placement.

For the control study, full memory partitioning is a context-side reference condition: it removes retained protected memory from the generation context. The PCMU-aware provenance gate is output-side: it retains the working context for internal computation but suppresses a candidate output unit when it matches protected PCMU information that is absent from the visible query under the same visible-query exclusion rule. The undefended wrapper retains 82.4% utility and exhibits 93.3% residual exposure. Full memory partitioning reduces exposure to 0% but retains only 59.5% utility, a 27.8% relative utility loss. The PCMU-aware provenance gate reduces residual exposure to 4%, retains 80.3% utility, and incurs a 2.5% relative utility loss.

G Ethics and Data Governance

The study uses previously released, de-identified clinical benchmark records and involves no participant recruitment, intervention, or live clinical decision making. We do not attempt re-identification. MIMIC-III (Johnson et al., 2016), MIMIC-IV (Johnson et al., 2023), MIMIC-CXR (Johnson et al., 2019), and eICU (Pollard et al., 2018) remain subject to the applicable access and data-use requirements; raw records and patient-linked traces are not redistributed. Because the benchmark

studies a privacy-relevant failure mode, the paper reports aggregate measurements, sanitized high-level strategies, and deterministic scoring logic while withholding patient-linked traces, credentials, private endpoints, and copy-pastable operational extraction payloads.

The benchmark does not claim to bypass production authentication, RBAC, hospital permissions, or physically isolated agent architectures. Clinical role framings are linguistic task contexts rather than authenticated identities. The methodological claim concerns verifier-guided probing-strategy acquisition and frozen memory-exposure assessment in the controlled benchmark, not the prevalence of any deployment architecture.

H Reproducibility, Code, and Data Availability

Artifact contents. The reproducibility artifact records the exact initial seed repertoire, structured-experience schema, data-split identifiers, candidate-generation manifests, deduplication and admission decisions, accepted, rejected, and pruned strategies, strategy/parent identifiers, supporting experience identifiers, generation, validation outcomes, admission status, text hashes, final repertoire and router checksums, model and decoding manifests, the PCMU normalizer, fixed random seeds, aggregate result tables, and plotting inputs. Appendix B reports implementation constants and rules from the evaluated code. Restricted MIMIC-family and eICU records are not redistributed; reproduction requires access through the corresponding official providers. Synthetic schema-compatible fixtures exercise strategy evolution, state classification, verification, aggregation, and plotting without restricted records. After double-blind review, we plan to release TRACE, configuration manifests, synthetic fixtures, and plotting scripts under an open-source license; restricted clinical records remain excluded.