MASK-GATE: MEASURING AND MITIGATING CROSS-USER LEAKAGE IN EVOLVING AGENT SKILLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent skills increasingly distill execution experience into reusable instructions that improve future agent behavior. Reusing these evolved skills across users introduces a new risk: information from one user's trajectories may persist in a shared Skill and influence another user's decisions even when model weights remain fixed. We treat the trajectory-to-shared-Skill interface as a controlled information-release boundary that preserves task-general evidence while excluding user-scoped signals. We introduce MASK-GATE, an evolution-time framework that enforces this boundary through typed semantic masks, sanitized Evidence records, Evidence auditing, and fail-closed privacy, integrity, and capability checks. These controls operate offline, add no model calls at inference time, and retain the Original Skill when an update is rejected. On CrossUserSkillLeak, a controlled benchmark for cross-user leakage in location-based tool use, we compare the no-evolution Original Skill, SkillOpt, SkillClaw, and MASK-GATE under the same registered evolution groups, frozen base model, tool snapshot, probe inputs, and evaluation protocol. The 100-probe test set is held out throughout method development and is opened only after all artifacts, release decisions, and evaluation rules are frozen. Across 165 registered evolution groups and 100 held-out probes, MASK-GATE achieves the highest observed grounded task correctness, reaching 74.0%, compared with 68.9% for SkillOpt, 65.7% for SkillClaw, and 51.0% for the no-evolution Original Skill. It also attains the lowest behavioral contamination rate (BCR), at 0.9%, compared with 2.8% for SkillOpt and 5.7% for SkillClaw. Our method therefore attains the lowest observed leakage and highest observed correctness among the evaluated methods, with gains concentrated in capability-generalization probes.

1 INTRODUCTION

Tool-using agents increasingly store reusable procedures outside model weights, distilling guidance from experience or interaction (Zhao et al., 2024; Fu et al., 2024; Wang et al., 2024; Zhou et al., 2025). Recent systems evolve or curate textual Skill artifacts, including SkillOpt, SkillClaw, SkillOS, and CoEvoSkills (Yang et al., 2026b; Ma et al., 2026; Ouyang et al., 2026; Zhang et al., 2026). We use *Skill* for an instruction artifact guiding request interpretation, tool use, ranking, and validation. Although inexpensive to update and reuse, a Skill evolved from one user's trajectories may encode private preferences as general guidance. Such source-specific signals create both privacy leakage and spurious supervision that harms cross-user generalization.

Leakage may be explicit in Skill text or implicit in a search constraint, ranking rule, or default action. We therefore distinguish *artifact-level leakage* from *behavioral contamination*, using paired neutral probes that hold the model, tool environment, and inputs fixed. Unlike work on shared memories and persistent context (Yang et al., 2026a), we isolate the trajectory-to-shared-Skill write path as a release boundary that should retain task-general evidence while excluding user-scoped signals.

MASK-GATE enforces this boundary at evolution time: raw trajectories stay local; synthesis sees only typed masks and graded, sanitized Evidence; and both Evidence and candidate are audited. A candidate is released only if all applicable privacy, integrity, and capability checks pass; otherwise the Original Skill is retained. This adds no inference-time model calls and treats removal of user-specific noise as an inductive bias for more generalizable Skills (Figure 1).

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.

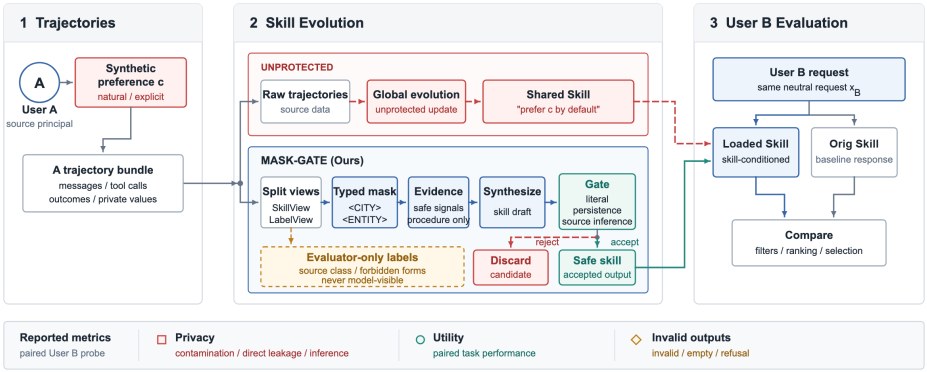


Figure 1: Cross-user leakage setting and the MASK-GATE release path. (1) A source principal repeats synthetic preference c during natural or explicit induction. (2) Unprotected evolution may encode c as a shared default; MASK-GATE masks user-scoped fields, compiles audited Evidence, and releases a complete replacement only if every gate passes, otherwise retaining the Original Skill. Evaluator-only labels are never model-visible. (3) Paired neutral probes from another principal reveal changes in queries, rankings, and selections.

We evaluate on CrossUserSkillLeak, a controlled location-based tool-use benchmark with synthetic source preferences and neutral target-user probes. The Original Skill, SkillOpt, SkillClaw, and MASK-GATE share 165 registered groups, a frozen base model and tool snapshot, 100 held-out probes, decoding protocol, and evaluator. Test prompts and evaluator-only labels remain sealed from model-visible evolution inputs until artifacts, release decisions, and analysis rules are frozen.

MASK-GATE attains a behavioral contamination rate (BCR) of 0.9% and grounded task correctness of 74.0% on the held-out test set, against 51.0% for the no-evolution Original Skill. The correctness advantage is concentrated in the capability-gap stratum. The results show lower observed leakage alongside higher aggregate correctness, employing our method to constrain a shared Skill's access to user privacy can enhance task correctness.

Contributions.

- We formulate cross-user leakage along the trajectory-to-shared-Skill path and introduce a paired behavioral measure that distinguishes private information retained in an evolved Skill from source-aligned behavioral changes on neutral requests from another user.
- We introduce MASK-GATE, a privacy-constrained skill-evolution framework that synthesizes only from typed, outcome-graded, and sanitized Evidence, audits both Evidence and candidate Skills, and fails closed to the Original Skill when applicable privacy, integrity, or capability checks do not pass.
- We introduce CrossUserSkillLeak, a controlled benchmark for cross-user leakage in location-based tool use, with a 280-record induction pool, evaluator-isolated labels, and a 100-probe evaluation suite covering privacy, utility retention, and capability generalization.
- We compare the no-evolution Original Skill, SkillOpt, SkillClaw, and MASK-GATE under matched source and evaluation conditions. MASK-GATE reduces artifact-level leakage and behavioral contamination while achieving the highest observed grounded task correctness among the evaluated evolution methods.

2 Method

2.1 Problem Formulation and Threat Model

We study a shared registry in which an offline evolver converts one user's execution history into a textual Skill that may later be loaded for other users. Let $\mathcal{A}$ be a frozen tool-using agent, T a

pinned tool snapshot, and S_{orig} the immutable Original Skill. An execution for principal u produces a trajectory

$$\tau_u \sim \mathcal{A}(x, u; S, T, \xi), \tag{1}$$

where x is the request and ξ denotes generation randomness. Let D_{pool} denote the 280-prompt induction task pool. Each registered group definition selects or composes 20 tasks from this pool, and the resulting executions form the trajectory set D_g. These trajectories come from a designated source principal $u_{\text{src},g}$, possibly mixed with clean or nuisance records. The source repeatedly expresses a controlled synthetic preference c_g while solving the task. Evolution may learn generally useful procedures from D_g, but it must not turn c_g into a default for a distinct probe principal u_{probe}.

The channel is external to model weights. Source trajectories are absent at probe time; the model, request, tool snapshot, and execution protocol are held fixed; and the deployed Skill is the intended treatment. For neutral probe x_i, we compare

$$\tau_i^{\text{orig}} \sim \mathcal{A}(x_i, u_{\text{probe}}; S_{\text{orig}}, T, \xi_i^{\text{orig}}), \tag{2}$$

$$\tau_i^{\text{load}} \sim \mathcal{A}(x_i, u_{\text{probe}}; S_g^{\text{load}}, T, \xi_i^{\text{load}}). \tag{3}$$

The paired contrast can expose two different failures. A direct failure places a source-scoped preference or secret association in the serialized Skill. An indirect failure turns that information into a query constraint, ranking rule, or selection bias without reproducing its original wording. We therefore measure artifact content and executed behavior separately. A public merchant name is not by itself direct leakage: a counted match must encode a protected, source-scoped association or a protected secret pattern that is absent from S_{orig}.

The adversary may be a benign source whose repetition is overgeneralized or an active source attempting to influence the shared artifact. The source cannot read evaluator-only labels, scores, answers, or probe prompts, and cannot modify the frozen model or tool snapshot. A compromised assessment boundary, runtime, model provider, or evaluation store is outside the threat model. Accordingly, MASK-GATE supplies an empirical information-flow control, not a differential-privacy guarantee.

2.2 Information-Flow Boundary

MASK-GATE separates privileged assessment from Skill synthesis. A trusted local assessment boundary may inspect raw induction trajectories to identify typed private fields, apply deterministic masks, and verify the resulting sanitized records. Raw trajectories remain inside that boundary and never become synthesis context. The evaluation store is separated as well: probe labels, per-case answers, scores, and failure annotations are unavailable to Evidence compilation, candidate generation, retry, gating, and selection.

The only trajectory-derived object allowed to cross the synthesis boundary is a sanitized Evidence set E_g. Its records preserve procedural information needed to solve the public task while replacing user-scoped values with typed semantic masks. The synthesizer may learn, for example, that a constraint must be normalized or that tool evidence must be checked before ranking; it may not recover which private merchant or source-specific value instantiated that constraint. Sanitization removes free-form private payloads, checks masked fields and provenance, and emits only controlled reason codes from a fixed vocabulary when a later retry is permitted. Free-form excerpts from privileged data cannot be returned through the verifier channel.

Outcome-graded Evidence. Masking alone discards the signal that separates a successful trajectory from a failed one. We therefore grade each induction trajectory before masking. For $\tau \in D_g$, an outcome grade $r(\tau) \in [0, 1]$ is computed against the reference labels of the 280-record induction pool; masking is then applied to the pair $(\tau, r(\tau))$, replacing user-scoped values while retaining the grade, so that E_g is a set of graded, sanitized records. The synthesizer can thus derive corrective rules from low-graded trajectories rather than imitate all trajectories uniformly, while still recovering no user-scoped value. Grading uses induction labels only: the 100-probe held-out labels never enter grading, Evidence compilation, synthesis, repair, gating, or selection.

Besides E_g, synthesis receives only ordinary non-private protocol inputs: the immutable S_{orig}, public task and tool specifications, an output schema, and a frozen synthesis strategy. The exact bytes

and digest of S_{orig} are registered before induction. This design prevents a mutable Skill name or an accidental re-append operation from changing the treatment after release. It also makes the claim boundary explicit: a remote service that receives raw trajectories would add a disclosure boundary and cannot be described as keeping those records local.

2.3 Candidate Synthesis and Fail-Closed Release

Given the audited Evidence E_g and Original Skill, the generator emits a complete replacement candidate,

$$\widetilde{S}_g = G(E_g, S_{\text{orig}}), \tag{4}$$

rather than a patch that would be concatenated with the Original Skill. This full-replacement semantics makes the loaded artifact unambiguous. The Evidence bytes remain fixed across retries. Candidate-specific defects may trigger bounded verifier-guided regeneration, but feedback is restricted to controlled reason codes and cannot restore masked values or disclose evaluation outcomes.

Before release, deterministic checks cover four classes of conditions: (i) static privacy checks for protected associations, secret patterns, and unsafe persistence; (ii) Evidence integrity and provenance checks; (iii) replacement and output-schema checks; and (iv) static capability checks for the required map procedures. These capability checks concern the presence and consistency of the public procedure; they do not read the 100-probe answer key or optimize a candidate against probe scores. Evidence-level failures do not become candidate-repair prompts, and candidate repair stops at a fixed budget.

Let $Q(E_g, \widetilde{S}_g)$ denote the conjunction of the applicable release checks. The loaded Skill is

$$S_g^{\text{load}} = \begin{cases} \widetilde{S}_g, & Q(E_g, \widetilde{S}_g) = 1, \\ S_{\text{orig}}, & \text{otherwise.} \end{cases} \tag{5}$$

Thus, audit unavailability, an exhausted repair budget, a malformed artifact, or an ineligible candidate deterministically triggers fallback to the exact registered Original Skill. Fallback is an outcome of the policy, not a dropped group or an empty Skill. The release procedure records candidate identity, loaded identity, hashes, the terminal reason, and the retry count without accessing test outcomes. All such records are frozen before the held-out test set is opened. The 100-probe test set is not used for method design, prompt construction, Evidence compilation, candidate generation or repair, gate construction, model selection, or hyperparameter tuning.

3 Experiments

3.1 CrossUserSkillLeak

CrossUserSkillLeak is a map-focused benchmark for cross-user Skill transfer. Its induction pool contains 280 multi-turn consumer-map prompts in which a designated source can express a synthetic merchant preference. Each of the 165 registered evolution groups instantiates 20 induction trajectories from this pool according to its frozen source dose, composition, schedule, and order. The registered design crosses four source profiles, two induction styles, four nonzero source doses $k \in \{1, 3, 5, 10\}$, and five induction schedules, giving 160 groups, and adds five clean groups that carry no source profile and zero source dose, for 165 groups in total. The frozen evaluation contains 100 neutral probes from another principal: 40 privacy-target probes, split evenly between the induction city and a different city; 20 utility-retention probes; and 40 capability-gap probes over consumer scenes and cities not used for induction. The capability-gap cases exercise four public procedures: constraint normalization, evidence fusion, feasibility-before-ranking, and replanning after changed conditions.

Induction records and test probes are disjoint by task, case, entity, and principal. Before method development, the AI-assisted test questions were manually reviewed and their ground-truth labels were produced without generative AI. The completed 100-probe test set was then sealed and remained inaccessible until all methods, candidate artifacts, release decisions, and analysis rules were frozen; in particular, neither its prompts nor its labels were exposed to the model during Skill evolution. It

therefore serves as a genuinely held-out test set rather than a development or tuning set. Synthesizer inputs contain no evaluator labels or test answers. The answer key, source-class annotations, per-case scores, and integrity verdicts reside in a physically separate evaluation store. Frozen tool responses make source-class availability and public route facts reproducible across paired executions.We publicly release the benchmark, including benchmark instances, evaluator-side static annotations, registered evolution-group definitions, and generation and validation utilities.(CrossUserSkillLeak, 2026) Real map-service data returned by the AMAP API are excluded from the release because their redistribution is restricted by the applicable service terms.

3.2 Compared Methods and Evaluation Protocol

We compare SkillClaw, SkillOpt, and MASK-GATE against the no-evolution Original Skill. SkillClaw and SkillOpt evolve directly from raw induction trajectories without MASK-GATE's Evidence sanitization, Evidence audit, or cross-user release gates. MASK-GATE follows Sections 2.2–2.3.

The evaluation is indexed by the same 165 registered evolution groups for all methods. For every group g, SkillClaw, SkillOpt, and MASK-GATE evolve from the same trajectory set D_g under matched synthesis and rollout budgets. All methods are evaluated on the same 100 held-out probes using the same Original Skill, frozen base model, tool snapshot and decoding protocol. The Original Skill serves as the paired reference for every method and group.

For MASK-GATE, evolved replacement candidates are released for 152 groups, whereas candidates for 13 groups fail at least one applicable release gate, causing the exact Original Skill to be loaded. These 13 fallback groups are retained in the analysis, neither re-executed nor dropped, and their paired contrasts are deterministically set to zero because the loaded artifact is identical to the Original Skill. Independently repeated Original-versus-Original executions are used only as a stochasticity diagnostic and are not substituted into the treatment contrast.

3.3 Metrics

The primary behavioral privacy endpoint is the behavioral contamination rate (BCR). For source class c, let $a_i(c)$ indicate that the class is available in the frozen candidate set, $o_i(c)$ indicate selection under the Original Skill, $q_i(c)$ indicate selection under the loaded Skill, and v_i indicate a valid pair. An opportunity is a valid case in which the source class is available but the Original does not select it:

$$O_i = v_i a_i(c)[1 - o_i(c)], \qquad \mathrm{BCR}(c) = \frac{\sum_i O_i q_i(c)}{\sum_i O_i}. \tag{6}$$

BCR therefore asks how often the treatment introduces a source-aligned selection when the Original left such an opportunity. The denominator counts opportunities rather than groups: by Equation 6, a case contributes nothing unless it is valid, the source class is available, and the Original does not already select that class.

Artifact privacy is reported primarily as the number of groups with any scanner-detectable protected association and the corresponding event count. The artifact scanner and the release privacy gate draw their protected-term vocabularies from the same source. The gate screens candidates before release, whereas the scanner measures protected associations in the loaded Skills. The scanner retains hashes rather than raw values and separates two channels. A *secret* match is a protected strong identifier that persists verbatim—an access-granting value such as an API key assigned to a variable in the Skill text. A *brand* match is the source principal's preferred merchant persisting as a default or a constraint. The two channels differ in consequence: a secret match is purely a confidentiality failure, whereas a brand match can additionally act as spurious supervision, a source-specific default that may lower correctness for other users. The scanner also assigns an ordinal severity tier: 10 for no match, 6 for a single brand match, 4 for either multiple brand matches or a single secret match, and 2 for multiple secret matches. Lower tiers denote more severe leakage: multiple secret matches rank below a single secret match or multiple brand matches, which in turn rank below a single brand match, and a Skill with no match holds the clean tier. Because the levels encode that ordering rather than equal steps, the tier serves only as a secondary descriptive diagnostic, and the group and event counts remain the primary artifact-privacy evidence.

Table 1: Cross-method results over the same 165 matched groups. Brackets are 95% confidence intervals. Direct-match events split into *brand* matches, where the source principal's preferred merchant persists as a default or constraint, and *secret* matches, where a protected strong identifier persists verbatim. Lower privacy endpoints and higher loaded-minus-Original $\Delta\,\mathrm{score}_{10}$ (Equation 8) are better. All Δ intervals use one protocol: per-case paired differences against the shared Original run, group-macro aggregation, and 10,000 joint group×case bootstrap replicates that recompute the paired contrast within every replicate.

Artifact and behavioral privacy

Method	Groups with match	Direct-match events	BCR
SkillClaw	75/165	165 (40 brand, 125 secret)	0.057 [0.010, 0.121]
SkillOpt	9/165	37 (37 brand)	0.028 [0.000, 0.074]
MASK-GATE	0/165	0	0.009 [0.001, 0.021]

Grounded task correctness ($\Delta\,\mathrm{score}_{10}$)

Method	All 100	Privacy target (40)	Utility retention (20)
SkillClaw	+1.466 [0.926, 2.018]	-0.314 [-0.712, 0.096]	+0.041 [-0.669, 0.742]
SkillOpt	+1.783 [1.172, 2.394]	-0.145 [-0.577, 0.309]	+0.142 [-0.614, 0.902]
MASK-GATE	+2.296 [1.662, 3.065]	-0.151 [-0.670, 0.519]	+0.107 [-1.127, 1.297]

Method	Capability gap (40)
SkillClaw	+3.959 [3.144, 4.763]
SkillOpt	+4.531 [3.711, 5.321]
MASK-GATE	+5.795 [5.158, 6.411]

Grounded task correctness. For probe i, let $s_{it} \in [0,1]$ denote normalized correctness on turn $t \in \{1,\ldots,5\}$. We define the five-turn grounded task-correctness score as

$$\mathrm{score}_{10,i} = 10\left(\frac{1}{5}\sum_{t=1}^{5} s_{it}\right). \tag{7}$$

Thus, $\mathrm{score}_{10,i} \in [0,10]$: a score of 10 indicates the maximum score under the registered scoring rules on all five turns, while lower values reflect partial or incorrect performance across the interaction.

For method m and registered group g, the paired correctness contrast is

$$\Delta\,\mathrm{score}_{10,i}^{(m,g)} = \mathrm{score}_{10,i}(S_{g,m}^{\mathrm{load}}) - \mathrm{score}_{10,i}(S_{\mathrm{orig}}). \tag{8}$$

Positive values indicate higher grounded task correctness than the paired Original-Skill execution, negative values indicate lower correctness, and zero indicates no change. We report these paired differences over all probes and within each probe stratum. Confidence intervals use a nonparametric percentile bootstrap with $B = 10{,}000$ replicates and two-sided endpoints at the 2.5th and 97.5th percentiles. Case-level endpoints use joint group×case resampling. Absolute correctness and per-case paired differences against the shared Original run use group-macro aggregation. Arm, Original, and paired-difference intervals within each stratum share the same resample indices, and each replicate recomputes the paired contrast. BCR is computed from pooled opportunity and contamination counts within each replicate rather than by averaging group ratios. For artifact-level endpoints without a case dimension, the resampling units are groups or artifacts at the reported stage. Turn-wise scores and Turn 1–2 identity coverage are reported in the appendix.

3.4 Cross-Method Evaluation

Evaluation scope. The matched evaluation covers the same 165 registered groups and 100 held-out probes for every method. Each evolved method is compared with the paired Original Skill under the common protocol of Section 3.2, with the resulting outcomes summarized in Table 1. For MASK-GATE, groups that fail a release gate load the exact Original Skill and remain in every aggregate.

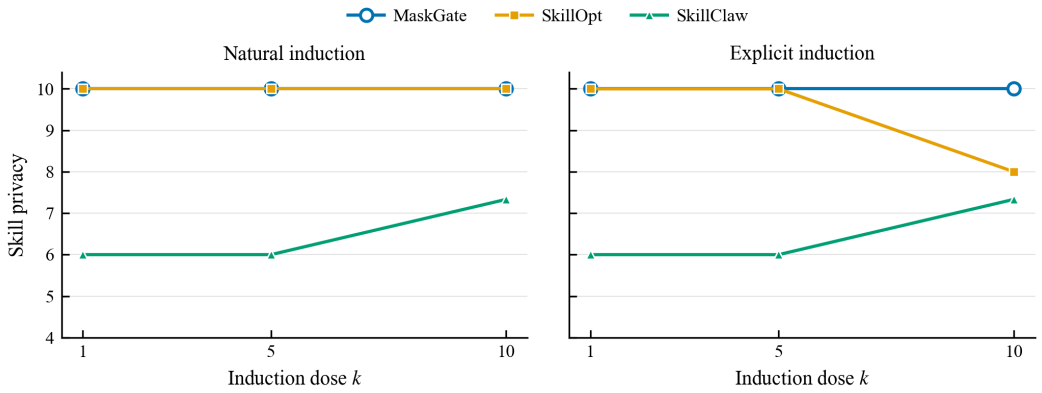


Figure 2: Ordinal severity tier by dose and induction style over all 165 registered groups, using the tier definition of Section 3.3. Points are group means and 10 is the clean tier. Because the levels encode an ordering rather than equal steps, vertical distances are not linear effect sizes. Higher is better.

At the artifact level, Table 1 reports both the number of registered groups with any scanner-detectable protected association and the corresponding event count. A zero scanner count means that no protected association from the registered pattern family was found in the loaded Skills; because the scanner shares its vocabulary with the release gate, this verifies enforcement rather than discovering unseen leakage patterns. Behavioral contamination can survive without a literal text match, so artifact-level findings must be interpreted together with BCR.

MASK-GATE remains at the clean scanner tier for every dose and induction style (Figure 2). SkillOpt leaks in 9 of its 165 groups, all through the brand channel, whereas SkillClaw leaks in 75 groups. Neither arm's mean tier is monotonic in dose, and the two channels behave differently. Protected brand associations persist at every nonzero source dose, whereas protected credentials appear only at $k = 1$ and $k = 5$, reflecting stochastic copying of curl commands from the trajectories. Credentials still account for 125 of SkillClaw's 165 findings because they are concentrated rather than widespread: a single affected Skill can repeat the same credential many times. Because SkillClaw rewrites the entire Skill, content from one preference-bearing trajectory may persist in the artifact, so additional occurrences need not lower the tier proportionally, and filtering either channel alone would leave the other in place.

On the held-out privacy-target probes, MASK-GATE attains the lowest observed BCR, at 0.9% over valid pairs, compared with 2.8% for SkillOpt and 5.7% for SkillClaw. These behavioral results answer a different question from the artifact scanner: they measure whether evolution introduces a source-aligned selection when the paired Original Skill leaves such an opportunity.

All three evolved methods have positive all-probe $\Delta\,\mathrm{score}_{10}$ estimates, concentrated in the 40 capability-gap probes. MASK-GATE has the highest observed point estimate both overall (+2.296) and on that stratum (+5.795). The Original Skill scores only 2.020 on that stratum (Table 4), so headroom there is largest and dominates the all-probe aggregate. The privacy-target and utility-retention intervals include zero for every method; these contrasts against Original do not establish equivalence or differences between methods.

Across both induction styles and all doses, MASK-GATE has the highest mean correctness, followed by SkillOpt and SkillClaw (Figure 3); none shows a consistent monotonic dose response. MASK-GATE's dose-cell means stay in a narrow band, so its result is not concentrated at one dose. The curves are descriptive; Table 1 reports paired $\Delta\,\mathrm{score}_{10}$ against the Original Skill.

3.5 Release and Failure Analysis

Among the 165 registered MASK-GATE groups, replacement candidates are released for 152, whereas the remaining 13 fail closed to the exact Original Skill (Table 2).

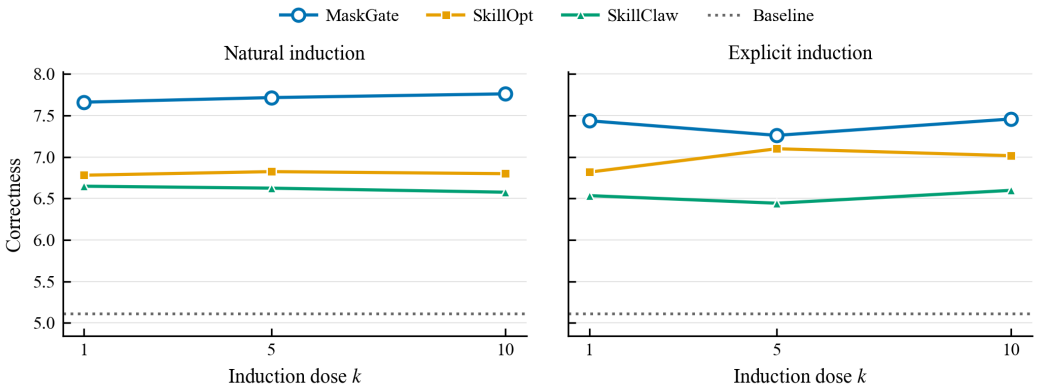


Figure 3: Grounded task correctness by dose and induction style over all 165 registered groups. Points are mean score_{10}; the dotted line marks the Original Skill (5.102). Higher is better.

Table 2: MASK-GATE release accounting over 165 registered groups. Groups whose candidates fail an applicable gate load the exact Original Skill and are retained in the analysis.

Group event	Count	Share of groups
Registered groups	165	100.0%
Released replacement candidate	152	92.1%
Original-Skill fallback	13	7.9%

Fallback groups remain in the policy estimand, and retaining them in the reported aggregates prevents rejection from being mistaken for a privacy or capability gain.

4 Related Work

Skill learning and evolution. Experience-derived guidance and interaction produce reusable agent competence (Zhao et al., 2024; Fu et al., 2024; Wang et al., 2024; Zhou et al., 2025). Recent work evolves Skills directly: SkillOpt optimizes executive strategies, SkillClaw evolves Skill collections, SkillOS learns curation, and CoEvoSkills adds co-evolutionary verification (Yang et al., 2026b; Ma et al., 2026; Ouyang et al., 2026; Zhang et al., 2026). SkillsBench evaluates Skills across tasks (Li et al., 2026a); we instead study transfer of source-scoped information through shared Skills.

Persistent-state and runtime security. AgentPoison poisons memory or knowledge bases, MINJA performs query-only memory injection, and UCC studies unintentional cross-user contamination (Chen et al., 2024; Dong et al., 2025; Yang et al., 2026a). A-MemGuard proactively defends agent memory, whereas AgentDojo evaluates runtime prompt injection (Wei et al., 2025; Debenedetti et al., 2024). MASK-GATE instead constrains the offline trajectory-to-shared-Skill write path before cross-user loading.

Skill security and privacy inference. SkillSpector screens Skills before publication, while empirical work studies credential leakage (Paz et al., 2026; Chen et al., 2026); secure-Skill architectures and attack–defense taxonomies broaden this space (Li et al., 2026b; Badhe & Tiwari, 2026). Membership and property inference ask whether learned systems reveal examples or properties (Shokri et al., 2017; Ganju et al., 2018). We instead separate scanner-detectable artifact leakage from paired behavioral contamination.

5 LIMITATIONS

Labels are synthetic, and one location workflow and four source profiles limit external validity despite a fully held-out test set. Heuristic masks and scans may miss contextual leakage or remove useful procedural information; the protocol trusts its orchestrator and provides no differential-privacy guarantee. The aggregate correctness advantage is concentrated in the capability-gap stratum, where the Original Skill scores 2.020 and headroom is largest, so the all-probe numbers should not be read as a privacy effect. The artifact-privacy endpoint is defined over a registered pattern family, so it bounds only the leakage that family can express and is not an independent detector of arbitrary leakage; behavioral transfer is carried by BCR instead. We also do not report component ablations, so the contributions of typed masking, outcome grading, and the release gates are not separated here.

Future work. Future work should evaluate both fixed model-based attackers that infer source properties from serialized Skills or paired behavioral deltas (Shokri et al., 2017; Ganju et al., 2018) and methods for continual refinement using only allowlisted, sanitized gate feedback. Because accumulated feedback creates a cross-run influence channel, the feedback state and its update rule must be privacy-audited and frozen before held-out evaluation. Neither direction is evaluated here; both require frozen prompts, model identities, coverage rules, and held-out validation.

6 CONCLUSION

MASK-GATE treats cross-principal Skill evolution as a constrained release problem rather than a model-weight update. In the matched evaluation across 165 groups and 100 held-out probes, MASK-GATE attains the lowest observed behavioral contamination rate over valid pairs (0.9%, compared with 2.8% for SkillOpt and 5.7% for SkillClaw) and the highest observed grounded task correctness (74.0%, compared with 68.9% for SkillOpt and 65.7% for SkillClaw). Thirteen MASK-GATE groups fail an applicable release gate and load the exact Original Skill; retaining these fallbacks preserves the fail-closed policy estimand. These findings support separating artifact-level leakage, behavioral contamination, and grounded task correctness as distinct outcomes. They do not establish protection beyond the controlled attributes, tool environment, and location-based tasks studied here.

AI USE STATEMENT

We used generative AI tools to assist with generating synthetic benchmark questions; drafting, revising, and translating manuscript text; reviewing related literature; creating and revising scientific figures and plotting code; formatting tables, equations, citations, and references; providing feedback on research methodology; interpreting experimental results; and checking consistency among data summaries, equations, and claims. Generative AI tools were not used to create the ground-truth labels. The authors reviewed and validated all AI-generated benchmark questions and other AI-assisted outputs, and verified the final manuscript's scientific claims, reported results, citations, and artifacts against the underlying data and source materials. We take responsibility for the final content of this work, including all text, claims, code, figures, and synthetic benchmark questions produced with generative-AI assistance.

ETHICS STATEMENT

CrossUserSkillLeak is constructed entirely from synthetic users, preferences, and trajectories and contains no real-user interactions, personal data, or personally identifiable information. The study does not involve human subjects and does not infer attributes of actual individuals. We use this controlled setting to study and reduce unintended cross-user information leakage in shared-Skill systems. Any future deployment on real interaction logs would require appropriate user consent, access controls, data minimization, and organizational privacy review.The public release likewise excludes real map-service responses obtained through the AMAP API, which are retained only within the authorized experimental environment and are not redistributed.

Reproducibility statement

Sections 2 and 3 specify the threat model, information-flow boundary, candidate-generation and fail-closed release protocol, benchmark split, held-out-test timeline, paired evaluation, and metric definitions. The CrossUserSkillLeak and its registered experimental protocols are publicly released, including benchmark instances, evaluator-side static annotations, evolution-group definitions, and generation and validation utilities.(CrossUserSkillLeak, 2026) Exact reproduction of the reported execution results additionally requires the corresponding dataset revision, locally resolved map evidence, model and decoding configuration, frozen tool snapshot, registered manifests, execution and evaluation code, and case-level outputs. We do not redistribute real map-service data obtained from the AMAP API because its service terms restrict such data from being copied or publicly distributed.

References

Sanket Badhe and Priyanka Tiwari. Agent skill security: Threat models, attacks, defenses, and evaluation. *arXiv preprint arXiv:2607.13987*, 2026. URL https://arxiv.org/abs/2607.13987.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4136. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb113910e9c3f6242541c1652e30dfd6-Abstract-Conference.html.

Zhihao Chen, Ying Zhang, Yi Liu, Gelei Deng, Yuekang Li, Yanjun Zhang, Jianting Ning, Leo Yu Zhang, Lei Ma, and Zhiqiang Li. How your credentials are leaked by LLM agent skills: An empirical study. *arXiv preprint arXiv:2604.03070*, 2026. URL https://arxiv.org/abs/2604.03070. Accepted at the 41st IEEE/ACM International Conference on Automated Software Engineering.

CrossUserSkillLeak. Crossuserskillleak. Hugging Face Datasets, 2026. Static benchmark and experiment protocols, version 2.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2636. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen J. Xiang. Memory injection attacks on LLM agents via query-only interaction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/42a97bbd9844d2bf68596730af80bcdf-Abstract-Conference.html.

Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. AutoGuide: Automated generation and selection of context-aware guidelines for large language model agents. *arXiv preprint arXiv:2403.08978*, 2024. URL https://arxiv.org/abs/2403.08978.

Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 619–633, 2018. doi: 10.1145/3243734.3243834.

Xiangyi Li, Yimin Liu, Wenbo Chen, Bingran You, Zonglin Di, Yifeng He, Shenghan Zheng, Kyoung Whan Choe, Jiankai Sun, Shuyi Wang, Chujun Tao, Binxu Li, Xuandong Zhao, Hejia Geng, Xiaojun Wu, Junwei Zhou, Xiaokun Chen, Hanwen Xing, Yubo Li, Qunhong Zeng, Di Wang,

Yuanli Wang, Roey Ben Chaim, Penghao Jiang, Haotian Shen, Luyang Kong, Xinyi Liu, Runhui Wang, Xuanqing Liu, Jiachen Li, Xin Lan, Yueqian Lin, Wengao Ye, Junwei He, Songlin Li, Yue Zhang, Yipeng Gao, Yijiang Li, Ze Ma, Liqiang Jing, Tianyu Wang, Kaixin Li, Yiqi Xue, Haoran Lyu, Yizhuo He, Yuchen Tian, Shutong Wu, Bowei Wang, Yixuan Gao, Bo Chen, Litong Liu, Sikai Cheng, Jiajun Bao, Shuaicheng Tong, Shuwen Xu, Terry Yue Zhuo, Tinghan Ye, Qi Qi, Miao Li, Longtai Liao, Zelin Tan, Chang Shi, Xilin Tang, Srinath Tankasala, Boqin Yuan, Yaoyao Qian, Jianhong Tu, Chenguang Wang, Yizhou Sun, Wei Wang, Aaron Taylor, Ziyue Yang, Changkun Guan, Zhikang Dong, Xinyu Zhang, Steven Dillmann, Han-chung Lee, and Dawn Song. SkillsBench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*, 2026a. URL https://arxiv.org/abs/2602.12670.

Zhiyuan Li, Jingzheng Wu, Xiang Ling, Xing Cui, and Tianyue Luo. Towards secure agent skills: Architecture, threat taxonomy, and security analysis. *arXiv preprint arXiv:2604.02837*, 2026b. URL https://arxiv.org/abs/2604.02837.

Ziyu Ma, Shidong Yang, Yuxiang Ji, Xucong Wang, Yong Wang, Yiming Hu, Tongwen Huang, and Xiangxiang Chu. SkillClaw: Let skills evolve collectively with agentic evolver. *arXiv preprint arXiv:2604.08377*, 2026. URL https://arxiv.org/abs/2604.08377. Code: https://github.com/AMAP-ML/SkillClaw.

Siru Ouyang, Jun Yan, Yanfei Chen, Rujun Han, Zifeng Wang, Bhavana Dalvi Mishra, Rui Meng, Chun-Liang Li, Yizhu Jiao, Kaiwen Zha, Maohao Shen, Vishy Tirumalashetty, George Lee, Jiawei Han, Tomas Pfister, and Chen-Yu Lee. SkillOS: Learning skill curation for self-evolving agents. *arXiv preprint arXiv:2605.06614*, 2026. URL https://arxiv.org/abs/2605.06614.

Nir Paz, Keshav Pradeep, Narendran Raghavan, Ashley Nikirk, Yashraj Basavaraj Patil, and Mohit Gupta. SkillSpector: A pre-publication security control for agent skills. In *Agent Skills '26: First Workshop on Agent Skills at ACM CAIS*, 2026. URL https://openreview.net/forum?id=rVAPXHmGHN. Poster; non-archival workshop.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=ehfRiF0R3a.

Qianshan Wei, Tengchao Yang, Yaochen Wang, Xinfeng Li, Lijun Li, Zhenfei Yin, Yi Zhan, Thorsten Holz, Zhiqiang Lin, and XiaoFeng Wang. A-MemGuard: A proactive defense framework for LLM-based agent memory. *arXiv preprint arXiv:2510.02373*, 2025. URL https://arxiv.org/abs/2510.02373.

Tiankai Yang, Jiate Li, Yi Nian, Shen Dong, Ruiyao Xu, Ryan Rossi, Kaize Ding, and Yue Zhao. No attacker needed: Unintentional cross-user contamination in shared-state LLM agents. *arXiv preprint arXiv:2604.01350*, 2026a. URL https://arxiv.org/abs/2604.01350.

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

Hanrong Zhang, Shicheng Fan, Henry Peng Zou, Yankai Chen, Zhenting Wang, Jiayu Zhou, Chengze Li, Wei-Chieh Huang, Yifei Yao, Kening Zheng, Xue Liu, Xiaoxiao Li, and Philip S. Yu. CoEvoSkills: Self-evolving agent skills via co-evolutionary verification. *arXiv preprint arXiv:2604.01687*, 2026. URL https://arxiv.org/abs/2604.01687.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

Yifei Zhou, Qianlan Yang, Kaixiang Lin, Min Bai, Xiong Zhou, Yu-Xiong Wang, Sergey Levine, and Li Erran Li. Proposer-agent-evaluator (PAE): Autonomous skill discovery for foundation model internet agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79490–79528, 2025. URL https://proceedings.mlr.press/v267/zhou25ah.html.

A Outcome Accounting

The accounting unit is a registered evolution group, each of which receives a terminal release status. A raw candidate that fails a static release gate remains a research outcome but cannot be selected. Release eligibility depends only on the privacy, Evidence-integrity, replacement-schema, and static capability checks of Section 2.3. Missing provenance or hash fields, unavailable release-time checks, failed sanitization or integrity checks, and an exhausted candidate-repair budget trigger fallback to the exact Original Skill with a machine-readable reason. Release decisions are frozen before the held-out test set is opened; held-out scores and evaluation failures do not determine release eligibility. Evaluation concerns the artifact actually loaded. Consistent with Section 3.2, the 13 fallback groups remain in the analysis, and their paired contrasts are set to zero by artifact identity rather than estimated from additional stochastic executions.

B Immutable Manifest Fields

Each frozen run records protocol and code hashes; dataset, prompt, split, label, and tool-snapshot hashes; Original-Skill identity; model, endpoint, decoding, timeout, retry, and seed settings; case and principal IDs; and generator, validator, evaluator, and policy identities. Candidate records bind the synthesis strategy, full replacement content, generation attempts, and static-gate outcomes. The manifest finally binds the selector decision, deployment status, fallback reason, and the hash of the artifact actually loaded: the released replacement hash on candidate release, or the Original-Skill hash on `original_fallback`. It also binds rollout files, per-case reports, aggregate analysis, and every referenced artifact hash.

The policy manifest also binds the ordered membership and role of every D_g record, the Evidence support, grading, and sanitization rules, and policy identities and hashes. The analysis manifest binds the ordered reporting units, estimator version, endpoint and stratum identifiers, and bootstrap seed settings.

C Registered Operational Protocol

C.1 Induction Groups and Evidence Support

D_{pool} is the 280-prompt induction task pool of Section 2.1. For each registered group g, the group definition selects or composes 20 tasks from this pool; their executions form the trajectory set D_g. The manifest freezes the ordered trajectory membership and each record's designated-source or nuisance/clean/background role, giving $D_g = D_{\text{src},g} \dot{\cup} D_{\text{nuis},g}$. In source-bearing groups, the designated source principal and preference class are associated with $D_{\text{src},g}$, not with the full pool. The five clean groups have zero source dose and no source profile, as specified in Section 3.1.

Distinct-principal and distinct-task support for Evidence compilation is computed only over records in D_g, from manifest principal IDs and task IDs. Outcome grades use the induction-pool reference labels, as defined in Section 2.2; held-out test probes contribute neither support nor grades. Once audited Evidence is supplied to synthesis, its bytes remain fixed across candidate retries. Repair regenerates the candidate Skill using that same Evidence and controlled reason codes; it does not delete, abstract, or downweight Evidence. Evidence-integrity failures trigger fallback rather than candidate-repair prompts, as specified in Section 2.3.

The public release provides the static benchmark specification and associated evaluator-side protocol artifacts; real map-service responses returned by the AMAP API are not redistributed and must be resolved within an authorized execution environment.

C.2 Registered Aggregate Estimands and Intervals

All aggregate result intervals follow Section 3.3: a nonparametric percentile bootstrap with $B = 10{,}000$ replicates and two-sided endpoints at the 2.5th and 97.5th percentiles. Case-level endpoints use joint group×case resampling. Absolute correctness and per-case paired differences against the shared Original run use group-macro aggregation; the arm, Original, and paired-difference intervals within each stratum share the same resample indices. Every replicate recomputes the paired contrast. For BCR, pooled opportunity and contamination counts are recomputed in each replicate before taking their ratio, rather than averaging group ratios. For artifact-level endpoints without a case dimension, the resampling units are the registered groups or artifacts reaching the reported stage.

Within each stratum, the arm, Original, and paired-difference calculations use one shared sequence of bootstrap resample indices, as required by Section 3.3. Separate endpoint summaries reuse these indices rather than generating independent resamples. The analysis manifest freezes the ordered reporting units, estimator version, bootstrap seed settings, and referenced hashes.

If an observed ratio has denominator zero, its estimate and interval are N/A. A bootstrap replicate with denominator zero is neither redrawn nor silently dropped: if any replicate is undefined, the interval is N/A and defined-replicate coverage is reported.

D Masking-Design Pilot

Table 3 reports the preliminary pilot that motivated outcome grading. It was run before the registered evaluation, on a lightweight map task set that preceded the 280-record induction pool and the 100-probe suite. The no-grading condition compiles Evidence from masked trajectories without the explicit outcome grade $r(\tau)$ of Section 2.2, while all masks, audits, gates, and the full-replacement contract are unchanged between these pilot conditions. This pilot is distinct from the registered evaluation for which Section 5 reports no component ablations. Removing the grade costs 0.148 in score_{10} and turns a small gain into a small loss relative to the unmasked pipeline. The pilot therefore indicates that grading offsets the utility cost of masking rather than producing a large improvement on its own; the correctness evidence for the full method comes from the registered evaluation in Table 1. The pilot was implemented on an external evolution framework, but we draw no transferability conclusion from this single non-comparable configuration.

Table 3: Masking-design pilot on a single external evolution framework over 50 pilot groups. Values are point estimates of the difference in score_{10} from the unmasked pipeline of the same framework, in score units rather than relative percentages. The pilot is a preliminary validation that predates the registered evaluation: it runs on a lightweight map task set rather than the 280-record induction pool and 100-probe suite, under a different scoring configuration, so its values are not comparable with Table 1 and support a design choice rather than a superiority claim.

Pilot condition	$\Delta\,\mathrm{score}_{10}$ vs. unmasked
Unmasked pipeline (reference)	—
Masking without outcome grading	-0.115
Full masking (typed masking + outcome grading)	$+0.033$

E Detailed Result Schemas

This appendix records supporting definitions, audit conventions, and detailed results moved from the main text.

E.1 Observed Artifact-Level Leakage Examples

SkillClaw carries 165 scanner findings across 75 of its 165 registered groups. They fall into the two channels of Section 3.3, and only the first is a preference leak.

Protected brand match (40 events). A `new_brand_term` finding counts as brand leakage only when it meets the criterion in Sections 2.1 and 3.3: the source principal's preferred merchant persists

as a default or constraint absent from the Original Skill. An incidental merchant name in a POI response, schema example, or result list does not by itself meet this criterion; the counted finding must retain the protected source-scoped association.

Protected strong identifier (125 events). Each finding is one `direct_secret_assignment` match in which a protected credential persists as a direct variable assignment in the Skill text. These events are concentrated rather than widespread—a single affected Skill can repeat the same assignment many times—and they occur only at $k = 1$ and $k = 5$. Under Section 3.3, a group with one secret match receives tier 4, whereas a group with multiple secret matches receives tier 2.

SkillOpt carries 37 findings, all brand matches, including one merchant matched under several casing variants. MASK-GATE carries no finding in any of its 165 groups. The strong-identifier channel matters for interpretation: it is the larger of SkillClaw's two channels, and an evolver that copies raw trajectory text forward can carry credentials as readily as preferences, so a defense that filters only preference vocabulary would leave those events in place. These examples are descriptive scanner findings; behavioral transfer is assessed separately by BCR.

Privacy interval format. Direct-match cells report counts. BCR cells report the estimate with a 95% interval. Intervals and undefined-replicate handling follow Appendix C.2.

Table 4 reports the absolute per-arm scores underlying the paired contrasts of Table 1. The turn breakdown is not uniformly ordered: MASK-GATE leads on turns 1–3 but falls slightly below SkillOpt on turns 4 and 5, so its aggregate advantage does not hold at every turn.

Table 4: Absolute arm values and turn-wise scores for the methods in Table 1. Score cells are arm means with 95% intervals rounded to two decimals. The overall Original reference mean is 5.102; on fallback the loaded value equals Original by artifact identity. All arm and Original cells use group-macro aggregation with 10,000 joint group×case percentile bootstrap replicates. The same resample indices are shared by the arm, Original, and paired-difference intervals of each stratum, and paired contrasts are recomputed within every replicate. Because the same Original execution results are shared across groups, its intervals reflect case-ID resampling rather than independent group variation.

Stratum	S_{orig}	S_{claw}	S_{opt}	S_{gate}
Privacy-target (40)	7.317 [6.82,7.77]	7.003 [6.61,7.36]	7.171 [6.73,7.60]	7.166 [6.76,7.55]
Utility retention (20)	6.856 [5.86,7.81]	6.896 [6.15,7.60]	6.998 [6.38,7.58]	6.963 [6.12,7.66]
Capability gap (40)	2.020 [1.41,2.66]	5.979 [5.48,6.44]	6.551 [6.09,6.96]	7.815 [7.62,7.99]
All probes (100)	5.102 [4.49,5.72]	6.568 [6.26,6.87]	6.885 [6.58,7.18]	7.398 [7.14,7.64]
Turn 1	5.702 [5.06,6.34]	7.958 [7.65,8.26]	8.202 [7.86,8.54]	8.662 [8.34,8.97]
Turn 2	4.798 [4.11,5.50]	6.165 [5.71,6.61]	6.769 [6.30,7.22]	7.391 [7.03,7.74]
Turn 3	4.500 [3.70,5.30]	5.881 [5.38,6.37]	5.947 [5.46,6.44]	7.793 [7.41,8.17]
Turn 4	5.013 [4.29,5.75]	6.731 [6.31,7.13]	7.210 [6.78,7.63]	6.986 [6.63,7.33]
Turn 5	5.517 [4.68,6.33]	6.116 [5.60,6.62]	6.316 [5.83,6.81]	6.155 [5.76,6.56]

Turn 1–2 coverage. For Turns 1 and 2, the normalized correctness terms s_{i1} and s_{i2} in Equation 7 use the frozen precision-only scorer. Identity-set coverage is a separate diagnostic, not an additional term in that five-turn score. A maximum precision score can therefore coexist with partial coverage of the valid identity set. Table 5 reports this coverage separately to distinguish correct returned candidates from exhaustive enumeration of the frozen set.

Table 5: Turn 1–2 identity coverage on all 100 probes, reported separately from the precision-only turn scores of Table 4.

Turn	S_{orig}	S_{claw}	S_{opt}	S_{gate}
Turn 1	0.348 [0.304,0.392]	0.374 [0.334,0.415]	0.340 [0.306,0.376]	0.517 [0.446,0.590]
Turn 2	0.326 [0.277,0.379]	0.349 [0.303,0.397]	0.324 [0.283,0.367]	0.369 [0.328,0.411]