BEYOND THE FINAL SKILL: REVALUING SKILL EVOLUTION HISTORY AFTER EXECUTOR REPLACEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Skill evolution produces a sequence of reusable Skill versions, but deployment usually keeps only the version selected by the source model. When that model is replaced, an earlier version may work better with the new executor. Existing transfer approaches typically reuse the source-selected Skill or optimize a new target-specific Skill, leaving this historical choice unexplored. We propose **HEIRS** (**H**istory-preserving **E**xecutor-aware **I**nheritance through **R**evaluation and **S**election), which preserves the Skill history, keeps a compact set of distinct candidates, and selects one unchanged archived Skill using separate target-side screening and confirmation sets. Across Qwen3.5 migrations and a DeepSeek-to-GLM cross-family transfer, HEIRS improves selected migrations: mean Qwen ALFWorld success rises by 4.5 percentage points over endpoint transfer, and DeepSeek-to-GLM Spreadsheet success rises by 11.5 points. Across Qwen Spreadsheet and ALFWorld migrations, HEIRS uses 52.4% fewer logical target evaluations than fresh SkillOpt while avoiding target-side Skill optimization. The code implementation is available at anonymous.4open.science/r/HEIRS.

1 INTRODUCTION

As LLM agents move from one-off tasks to long-term deployment, they need to retain and reuse what they have learned. Skill-evolution methods address this need by using execution feedback to improve reusable task-solving procedures over time (Yang et al., 2026; Liu et al., 2026). This process leaves a sequence of Skill versions and selects one for deployment under the current language model. The model may later be replaced because of an upgrade, a provider change, or deployment constraints.

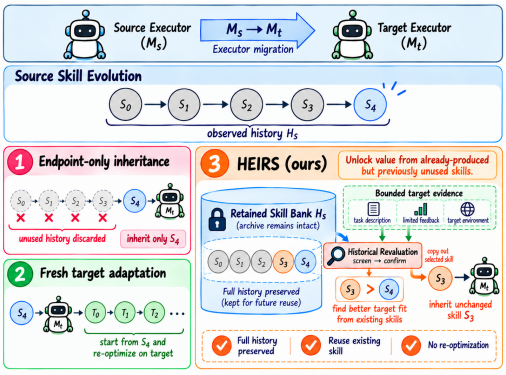

Figure 1: **Three choices after executor replacement.** Reuse the endpoint, optimize a new Skill, or revalue history with HEIRS.

A Skill that works well with one model may not work equally well with another (Tang et al., 2026; Huang et al., 2026; Yu et al., 2026); optimized prompts show the same sensitivity (Wang et al., 2026b). After migration, a system can reuse the source-selected Skill, which we call the endpoint, or adapt a Skill for the target model. A third possibility is usually ignored: an earlier version may already fit the target model better. We therefore ask: *which archived Skill should the target model inherit?*

Our experiments show that migration can reorder Skills within the same history. On a shared 16-task Spreadsheet panel, the source-selected endpoint solves six tasks under the target model, while a source-rejected version solves eleven. Source validation does not reveal this difference: Skills with the same source score can succeed on different tasks, and in two examined failures the program computes the correct answer but does not place it in the evaluated workbook. Thus a Skill rejected during source evolution can still be a better choice after migration. This motivates re-evaluating the preserved history before spending resources on new optimization (Figure 1).

We therefore propose **HEIRS** (**H**istory-preserving **E**xecutor-aware **I**nheritance through **R**evaluation and **S**election). HEIRS preserves the source history, keeps a compact set of distinct candidates, screens

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.

them on shared target tasks, and confirms a shortlist on a separate target set before transferring one archived Skill unchanged. It searches the archive instead of creating a new target-specific Skill.

We evaluate HEIRS across six directed migrations among Qwen3.5-4B/9B/27B, a DeepSeek-to-GLM cross-family migration, and Qwen9-to-Llama8 transfer. Compared with endpoint transfer, HEIRS improves mean Qwen ALFWorld success by 4.5 percentage points and DeepSeek-to-GLM Spreadsheet success by 11.5 points. The benefit is not universal: on Qwen9→Llama8 Spreadsheet transfer, HEIRS matches NoSkill and trails Direct Endpoint. Across Qwen Spreadsheet and ALFWorld migrations, HEIRS also uses 52.4% fewer logical target evaluations than fresh SkillOpt. Together, these results show that historical revaluation can recover useful prior work, but its benefit depends on both the alternatives preserved in history and the ability to identify the right one from limited target evidence.

Our contributions are:

- We identify **historical option value** after migration: source-rejected Skills can outperform the endpoint, with reordering linked to task-level gains, losses, and a concrete output failure.
- We propose **HEIRS**, which uses source-only history coverage, target screening, and disjoint confirmation to select an unchanged archived Skill under a fixed evaluation budget.
- We evaluate **benefits, selection losses, and costs** across model families and task domains, showing when historical revaluation helps and when limited target evidence causes selection errors.

2 Background

2.1 Preliminaries

A Skill S is a persistent natural-language procedure executed by an underlying language model M, which we call the *executor*. Under a source executor M_s, Skill evolution produces a history $\mathcal{H}_s = \{S^{(0)}, \ldots, S^{(L-1)}\}$ of L accepted and rejected versions, and selects an endpoint $S_{\text{end}} \in \mathcal{H}_s$ for deployment. The endpoint is source-selected, not necessarily the last generated version.

For a task set D, we define a Skill's empirical utility as

$$U(S; M, D) = \frac{1}{|D|} \sum_{x \in D} R(S, M, x), \qquad R(S, M, x) \in [0, 1], \tag{1}$$

where $R(S, M, x)$ is the benchmark evaluator's score; for binary outcomes, U is the fraction of tasks solved. At migration, M_s is replaced by a target executor M_t, while the task domain and evaluator remain fixed. Throughout the paper, $A \to B$ denotes migration of a Skill produced under source executor A to target executor B. Direct inheritance transfers S_{end}, whereas historical selection uses limited target-side evaluations to choose $\widehat{S} \in \mathcal{H}_s$. Historical selection changes neither the Skill text nor model parameters; selection and final test sets are disjoint.

2.2 Persistent Skill Evolution

Prior work allows agents to retain reusable experience as executable skills, linguistic feedback, or induced workflows (Wang et al., 2024; Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2025). More recent methods directly optimize persistent natural-language Skills. SkillOpt keeps the agent model fixed and iteratively updates the Skill from scored execution trajectories, accepting changes only when they improve held-out validation performance (Yang et al., 2026). RethinkSkill studies feedback dynamics across repeated Skill updates, while Trace2Skill distills trajectory-level lessons into reusable Skills (Liu et al., 2026; Ni et al., 2026). These methods create or improve source-model Skills; HEIRS instead asks which historical version should survive migration.

2.3 Cross-Model Portability and Adaptation

SkillLens shows that the same Skill can have different effects across target models (Huang et al., 2026). Prompt-transfer studies (Su et al., 2022) similarly report cross-model variation for optimized prompts. Existing methods respond to this mismatch in several ways. MASA adapts an inherited Skill to the target backbone through model-aware evolution and rewriting, while PromptBridge transfers

optimized prompts through source-to-target prompt mapping (Yu et al., 2026; Wang et al., 2026b). Other methods construct new target-conditioned solutions: SkillTTA retrieves relevant trajectories and synthesizes a temporary Skill, OPRO searches for improved prompts through iterative evaluation, and MaAS selects query-dependent agent structures from an agentic supernet (Wang et al., 2026a; Yang et al., 2024; Zhang et al., 2025). WikiSkill accumulates execution experience in a persistent wiki to support Skill evolution (Tang et al., 2026). HEIRS instead asks whether the source history already contains a better Skill for the target model.

2.4 Historical Options under Limited Evaluation

Reusing historical alternatives is well established in other settings: Snapshot Ensembles reuse training checkpoints, Model Soups combine multiple fine-tuned models, and checkpoint selection has been used to improve adversarial transferability (Huang et al., 2017; Wortsman et al., 2022; Li et al., 2025a). Historical Skills differ in two important ways: they are discrete natural-language artifacts, and model migration can change which version performs best. Selecting among these versions with limited target-side evidence is closely related to fixed-budget best-arm identification and sample-efficient evaluation (Shi et al., 2024; Kossen et al., 2021; Vivek et al., 2024; Maia Polo et al., 2024; Li et al., 2025b). HEIRS uses source-only information for shortlisting and shared target tasks for ranking.

3 Why Skill History Matters after Executor Replacement

The Skill selected by the source model is not necessarily the best choice after migration. We first show this reordering, then examine its task-level effects and one concrete execution mechanism.

3.1 Historical Rankings Change with the Executor

We evaluate 13 versions from a 4B SkillOpt history on the same 16 Spreadsheet tasks under 4B, 9B, and 27B executors. Skill texts and task identities remain fixed; utility is defined in Equation 1.

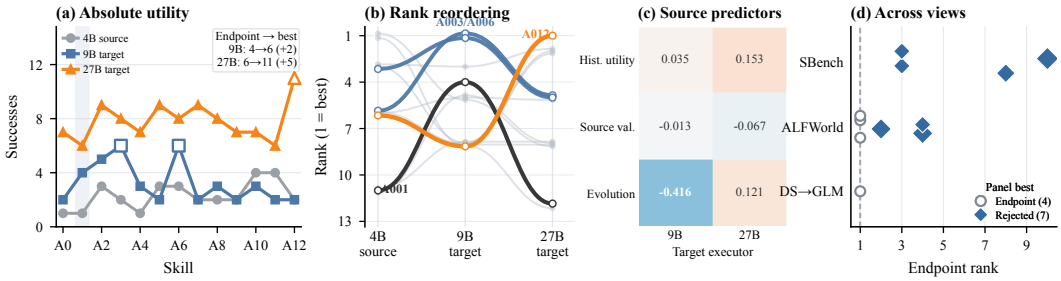

Figure 2: **Executor replacement changes utility and ranking within the same history.** (a) Successes and (b) ranks of 13 Skill versions on 16 shared tasks; tied ranks are offset. (c) Spearman correlations between source-side signals and target utility. (d) Endpoint ranks across 11 complete Spreadsheet and ALFWorld views, with DeepSeek→GLM shown separately. Marker shapes distinguish endpoint winners from source-rejected winners, and marker area shows additional successes over the endpoint (0–4). No accepted intermediate is best on any panel.

The ranking changes substantially after executor replacement. In all six directed Qwen views, at least one archived Skill solves more of the shared 16 tasks than the source endpoint, with gains of 1–5 tasks (Figure 2). Across 11 complete Spreadsheet and ALFWorld views, the endpoint is not best in seven; in every such case, the best Skill had been rejected during source evolution. Source validation score and checkpoint position do not reliably identify these target-preferred versions. The target executors also disagree with each other: over the same 13 Skills, the 9B↔27B Spearman rank correlation is only $\rho = -0.034$. The migration problem is therefore a selection problem over the history, not simply a question of keeping the latest or source-selected version. Because some views share a source history, they are descriptive rather than independent replications (Appendix A.1).

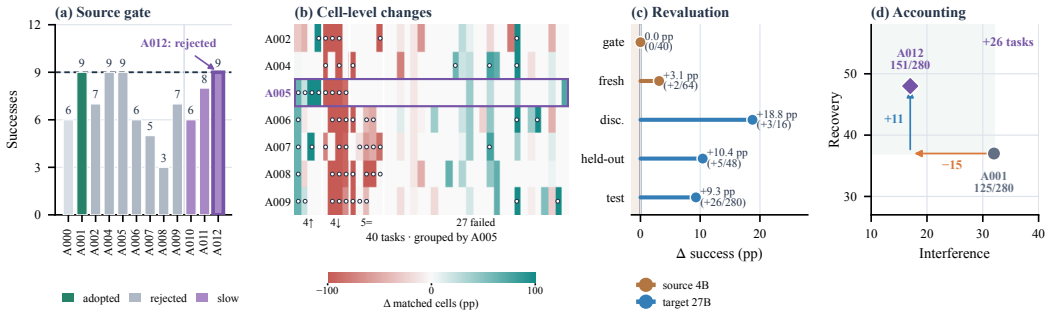

Figure 3: **Source rejection can hide useful task-level changes.** (a) The 4B source-validation gate selects only A001. (b) Matched-cell accuracy changes of seven rejected proposals relative to A001 across 40 tasks; dots indicate task success changes. For A005, four tasks improve, four regress, five remain successful, and 27 fail under both versions. (c) A012–A001 performance differences across the source gate and separate target panels. (d) On the same 280-task 27B test set, A012 gains 11 additional recoveries and avoids 15 disruptions, yielding 26 more successes than A001.

For a target task set D_t, define the *historical option value* $V_{\text{hist}}(D_t)$ as

$$V_{\text{hist}}(D_t) = \max_{S \in \mathcal{H}_s} U(S; M_t, D_t) - U(S_{\text{end}}; M_t, D_t) \geq 0. \tag{2}$$

For 4B→27B, the source-rejected A012 solves $11/16$ tasks, compared with $6/16$ for the endpoint, giving $V_{\text{hist}} = 5/16$. This quantity is the gain available from choosing the best archived Skill instead of the endpoint under a limited target-side evaluation budget.

3.2 Aggregate Scores Hide Different Task-Level Effects

Historical versions with similar source scores can behave differently at the task level. A001 and the rejected A005 both solve $9/40$ source tasks, yet each succeeds on four tasks that the other fails. Even among their 27 shared failures, A005 improves the fraction of correctly matched workbook cells by 6.56 percentage points. Across seven rejected fast-update proposals, partial outcomes include 30 improvements and 29 regressions (Figure 3a,b), so source score alone can hide task-level effects.

The same source score can therefore hide very different effects after migration. A012 receives the same $9/40$ source score as A001 but performs better on several target panels. To show where the difference comes from, we compare each Skill with NoSkill on the same tasks. For executor m, let p_m be the NoSkill success rate, $r_m(S)$ the fraction of NoSkill failures that Skill S turns into successes, and $h_m(S)$ the fraction of NoSkill successes that S turns into failures. Conditioning on the NoSkill outcome gives the resulting success rate $\mu_m(S)$:

$$\mu_m(S) = \underbrace{p_m}_{\text{native success}} + \underbrace{(1-p_m)r_m(S)}_{\text{recovered failures}} - \underbrace{p_m h_m(S)}_{\text{disrupted successes}}. \tag{3}$$

On the 280-task 27B panel, A012 solves 151 tasks compared with 125 for A001. Its 26-task advantage comes from 11 additional recoveries and 15 fewer disruptions (Figure 3d). Historical Skills therefore differ in both overall accuracy and which failures they fix or successes they break; this decomposition is descriptive rather than causal (Appendices A.2–A.3).

3.3 A Correct Answer Can Be Lost at the Output Step

The task-level differences above can arise from how an executor follows the same Skill instructions. On the same 128 Spreadsheet tasks, replacing A001 with A012 raises 27B successes from 56 to 71, but lowers 9B successes from 40 to 34. A012 extends A001 with corrective instructions, including computing answers in Python and writing their values to the workbook. We next inspect the executions to identify one concrete way in which the same historical guidance can lead to different outcomes.

An *uncached formula* stores an Excel expression without its computed value, causing the value-based evaluator to read an empty cell instead of the intended answer. Among 124 tasks with complete

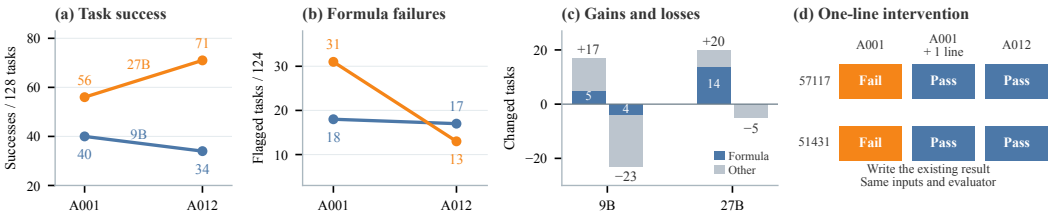

Figure 4: **A computed answer may not reach the evaluated workbook.** (a) A001 and A012 successes on the same 128 tasks. (b) First-error cells containing an uncached formula among 124 tasks with complete artifacts. (c) Gains and losses between A001 and A012; blue marks this failure signature in the losing version. (d) Two fixed 27B programs pass after changing only one assignment.

artifacts, this failure signature decreases from 31 to 13 cases under 27B, but only from 18 to 17 under 9B. It appears in A001 for 14 of the 20 tasks newly solved by A012 under 27B, compared with five of the 17 newly solved under 9B. This pattern is not exhaustive; some formulas also contain logic errors.

We test the output step directly with two fixed 27B programs. Both programs already compute the correct answer in Python but write an uncached formula to the workbook. Replacing only the final assignment with the existing result variable makes both programs pass under the same inputs and evaluator (Figure 4d). No computation, model call, or input is changed. For these two programs, the failure lies in delivering an already-computed answer, not in computing it. The intervention does not identify the responsible Skill instruction; it shows that identical historical guidance can produce different execution behavior under another executor.

4 HEIRS: History-Preserving Executor-Aware Inheritance

Section 3 shows that the best Skill can change after executor replacement. HEIRS uses a limited target-evaluation budget to find a better historical Skill while addressing two distinct errors: removing a useful candidate too early and choosing the wrong Skill among the retained candidates. As shown in Figure 5, it follows four stages: *Preserve*, *Admit*, *Revalue*, and *Commit*.

4.1 From Available Value to Selected Gain

HEIRS selects from the existing history rather than optimizing a new Skill. Let $\mathcal{C} \subseteq \mathcal{H}_s$ be the admitted set containing S_{end}, $\mathcal{C}_{\text{short}} \subseteq \mathcal{C}$ its shortlist, and $\widehat{S} \in \mathcal{C}_{\text{short}}$ the final choice. Let $R_{\text{cand}}(D)$ and $R_{\text{sel}}(D)$ denote admission and selection loss on a target set D, respectively. Then

$$\underbrace{U(\widehat{S}; M_t, D) - U(S_{\text{end}}; M_t, D)}_{\text{selected gain}} = \underbrace{V_{\text{hist}}(D)}_{\text{available value}} - \underbrace{R_{\text{cand}}(D)}_{\text{admission loss}} - \underbrace{R_{\text{sel}}(D)}_{\text{selection loss}}. \tag{4}$$

Admission loss measures the value lost when a useful Skill is excluded before target evaluation. Selection loss measures the value lost when the retained Skills are ranked incorrectly. We further write $R_{\text{sel}} = R_{\text{screen}} + R_{\text{conf}}$ to distinguish errors made during screening from errors made during final confirmation. As formalized in Appendices B.1–B.2, this decomposition separates two questions: which Skills deserve target evaluation, and which retained Skill should be transferred.

4.2 Preserve and Admit: Cover Historical Alternatives

Preserve keeps every valid, distinct Skill produced during source evolution, including rejected proposals, while removing exact duplicates and malformed artifacts.

Admit chooses a smaller set for target evaluation. History-Cover uses textual distance only to avoid spending the budget on redundant Skills; it does not use text to predict which Skill will perform best on the target executor. Let $\phi(S)$ denote the textual representation of Skill S. For Skill versions S_i and S_j, define their distance as

$$d(S_i, S_j) = 1 - \cos\big(\phi(S_i), \phi(S_j)\big). \tag{5}$$

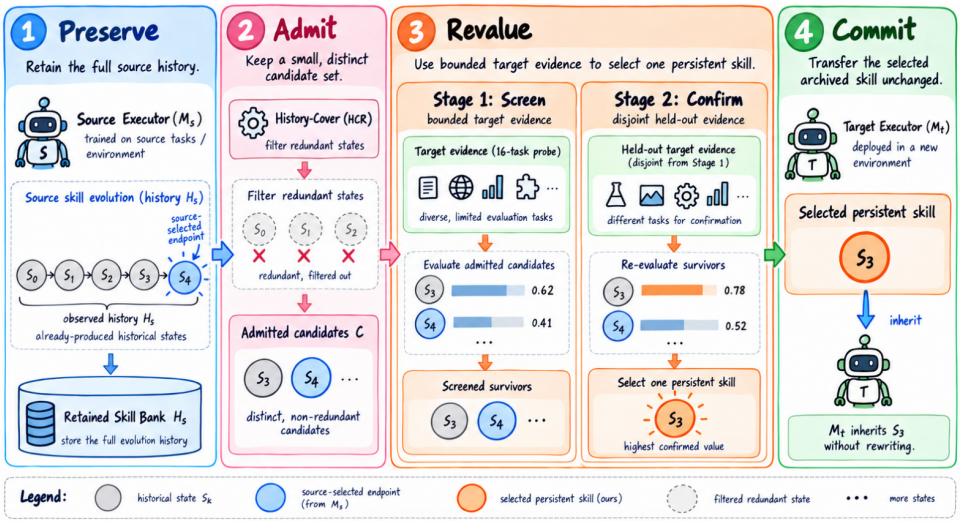

Figure 5: **HEIRS in four stages.** *Preserve* retains the source Skill history. *Admit* keeps distinct candidates using source-only information. *Revalue* screens the retained candidates and confirms the shortlist on disjoint target tasks. *Commit* transfers one archived Skill unchanged.

Let q control the granularity of historical coverage. History-Cover retains representative and boundary Skills based on this distance and always includes the source endpoint:

$$\mathcal{C} = \mathrm{HCover}(\mathcal{H}_s; q), \qquad S_{\mathrm{end}} \in \mathcal{C}. \tag{6}$$

Source-side content determines retention, while target-side evidence determines ranking; implementation details and admission limits appear in Appendices B.3–B.4.

4.3 Revalue: Screen and Confirm

Revalue first compares all admitted Skills on a shared screening set and then evaluates only the shortlist on new target tasks. On a common target set D_{scr}, each admitted Skill receives the screening score $\widehat{U}_{\mathrm{scr}}(S)$:

$$\widehat{U}_{\mathrm{scr}}(S) = \frac{1}{|D_{\mathrm{scr}}|} \sum_{x \in D_{\mathrm{scr}}} R(S, M_t, x). \tag{7}$$

All candidates are scored on the same target tasks for paired comparison, and a fixed shortlist rule retains the leaders in $\mathcal{C}_{\mathrm{short}}$. Screening narrows the set but does not choose the winner.

The shortlisted Skills are then evaluated on a disjoint confirmation set D_{conf}. HEIRS selects

$$\widehat{S} \in \arg\max_{S \in \mathcal{C}_{\mathrm{short}}} \frac{1}{|D_{\mathrm{conf}}|} \sum_{x \in D_{\mathrm{conf}}} R(S, M_t, x), \qquad D_{\mathrm{scr}} \cap D_{\mathrm{conf}} = \varnothing. \tag{8}$$

Only the confirmation results determine the final choice; screening scores are used only to form the shortlist. If the leaders remain tied, HEIRS evaluates the tied candidates on additional predefined confirmation tasks up to a fixed budget cap (Appendices B.5–B.6). Fresh confirmation therefore reduces the risk of overfitting to the initial screening set.

4.4 Commit: Unchanged Transfer under a Budget

Commit transfers $\widehat{S}$ unchanged, without rewriting, merging, or further optimization. Held-out tasks are used only for final evaluation. For migration i, let $K_i = |\mathcal{C}^{(i)}|$ and $L_i = |\mathcal{H}_s^{(i)}|$ be the number of admitted candidates and the archived history size. Let $B_{\mathrm{scr}} = |D_{\mathrm{scr}}|$ be the screening budget per candidate and $T_{\mathrm{conf}}^{(i)}$ the confirmation cost. We summarize the total target-side selection cost T_i and

fractional screening-cost saving $\eta_{\mathrm{scr}}^{(i)}$ as

$$T_i = B_{\mathrm{scr}} K_i + T_{\mathrm{conf}}^{(i)}, \qquad \eta_{\mathrm{scr}}^{(i)} = 1 - \frac{B_{\mathrm{scr}} K_i}{B_{\mathrm{scr}} L_i} = 1 - \frac{K_i}{L_i}. \tag{9}$$

One logical target evaluation executes one Skill on one target task.

5 Experiments

5.1 Experimental Setup

We evaluate transfer performance, historical-value loss, and target-side cost.

Benchmarks and transfer settings. Experiments cover SpreadsheetBench-Verified (Ma et al., 2024) and ALFWorld (Shridhar et al., 2021) across all six directed Qwen3.5-4B/9B/27B pairs, one cross-family migration from DeepSeek-V4-Flash to GLM-5.3-Flash, and Qwen9→Llama8 Spreadsheet transfer. Held-out Qwen sets contain 128 Spreadsheet and 48 ALFWorld tasks; cross-family Spreadsheet tests use 52 tasks per migration (Appendices C.1–C.2).

Baselines. Direct Endpoint transfers the source-selected Skill unchanged; matched Trace2Skill, EvoSkill, and WikiSkill implementations likewise transfer one source-selected artifact without target-side updates (Ni et al., 2026; Alzubi et al., 2026; Tang et al., 2026). Full History evaluates every archived version on the same screening set and performs one-stage selection, serving as an exhaustive archive-selection reference rather than a held-out oracle or equal-budget HEIRS variant. For admission analysis, Source Top-K_i uses source-validation scores to retain the same number of candidates as History-Cover. NoSkill and the released Human Skill provide reference performance. Target-side construction and adaptation baselines include Target Scratch SkillOpt, MASA, and SkillTTA (Yang et al., 2026; Yu et al., 2026; Wang et al., 2026a). Scratch uses the official SkillOpt engine, native train/validation split, and validation gate (Appendix C.3).

Selection and evaluation protocol. History-Cover uses $q = .45$, with $B_{\mathrm{scr}} = 16$ screening tasks and an initial $B_{\mathrm{conf}} = 16$ confirmation tasks. Ties use predefined benchmark-specific extensions. All operating points are chosen on separate development data before held-out testing. Screening, confirmation, and final testing use disjoint task sets, and every method fixes its Skill before final evaluation. Section 3 uses separate diagnostic evidence and does not affect the held-out evaluation.

5.2 When Historical Revaluation Improves Transfer

Table 1 reports held-out transfer performance after each method fixes its Skill. Appendix D.1 provides the target-side construction results not shown in the main table.

Historical revaluation helps on some migrations, not all. On ALFWorld, HEIRS raises the six-direction mean from 54.5% to 59.0%, improving three directions and matching Direct Endpoint on the other three. On Spreadsheet, HEIRS improves both migrations from 4B: 43.8% to 50.0% when moving to 27B and 31.3% to 32.0% when moving to 9B.

Fresh confirmation can correct the screening choice. On DeepSeek-to-GLM Spreadsheet, screening favors A007, but the separate confirmation set selects A009, which reaches 71.2% compared with 59.6% for Direct Endpoint. On ALFWorld, confirmation returns the endpoint at 87.5% rather than the 83.3% screening winner, again showing the value of fresh confirmation.

Qwen9-to-Llama8 Spreadsheet transfer provides a clear failure case. On Qwen9→Llama8 Spreadsheet transfer, HEIRS matches NoSkill at $4/52$, while Direct Endpoint reaches $5/52$. Historical revaluation therefore helps only when the archive contains a better target-compatible option and the evidence identifies it (Appendix D.4).

5.3 Where HEIRS Retains and Loses Historical Value

Equation 4 separates admission loss from selection loss; Figure 6 examines both stages.

Table 1: **Held-out Skill transfer after executor replacement.** Success (%) and percentage-point gains over matched NoSkill; bold marks HEIRS gains over Direct Endpoint. Qwen Avg. is the six-direction within-Qwen mean; Cross Avg. averages DS→GLM and Q9→L8 and is shown when both are available.

	Within Qwen							Cross-family		
Method	9→4	27→4	4→9	27→9	4→27	9→27	Qwen Avg.	DS→GLM	Q9→L8	Cross Avg.
SpreadsheetBench-Verified										
NoSkill	14.1		25.8		39.8		26.6	46.2	7.7	26.9
Human Skill	14.1 (+0.0)		21.9 (-3.9)		41.4 (+1.6)		25.8 (-0.8)	38.5 (-7.7)	7.7 (+0.0)	23.1 (-3.8)
Trace2Skill	17.2 (+3.1)	14.8 (+0.8)	26.6 (+0.8)	26.6 (+0.8)	41.4 (+1.6)	43.0 (+3.1)	28.3 (+1.7)	28.8 (-17.3)	15.4 (+7.7)	22.1 (-4.8)
EvoSkill	18.8 (+4.7)	14.1 (+0.0)	21.9 (-3.9)	21.9 (-3.9)	47.7 (+7.8)	42.2 (+2.3)	27.7 (+1.2)	32.7 (-13.5)	5.8 (-1.9)	19.2 (-7.7)
Direct Endpoint	20.3 (+6.3)	20.3 (+6.3)	31.3 (+5.5)	28.9 (+3.1)	43.8 (+3.9)	57.0 (+17.2)	33.6 (+7.0)	59.6 (+13.5)	9.6 (+1.9)	34.6 (+7.7)
WikiSkill	14.8 (+0.8)	14.1 (+0.0)	24.2 (-1.6)	21.9 (-3.9)	36.7 (-3.1)	38.3 (-1.6)	25.0 (-1.6)	26.9 (-19.2)	9.6 (+1.9)	18.3 (-8.7)
HEIRS (Ours)	19.5 (+5.5)	19.5 (+5.5)	**32.0** (+6.3)	24.2 (-1.6)	**50.0** (+10.2)	57.0 (+17.2)	**33.7** (+7.2)	**71.2** (+25.0)	7.7 (+0.0)	**39.4** (+12.5)
Full History	20.3 (+6.3)	15.6 (+1.6)	29.7 (+3.9)	30.5 (+4.7)	55.5 (+15.6)	58.6 (+18.8)	35.0 (+8.5)	59.6 (+13.5)	9.6 (+1.9)	34.6 (+7.7)
Scratch SkillOpt	20.3 (+6.3)		27.3 (+1.6)		48.4 (+8.6)		32.0 (+5.5)	65.4 (+19.2)	7.7 (+0.0)	36.5 (+9.6)
ALFWorld										
NoSkill	25.0		27.1		25.0		25.7	83.3	16.7	50.0
Human Skill	45.8 (+20.8)		47.9 (+20.8)		52.1 (+27.1)		48.6 (+22.9)	87.5 (+4.2)	16.7 (+0.0)	52.1 (+2.1)
Trace2Skill	47.9 (+22.9)	52.1 (+27.1)	39.6 (+12.5)	33.3 (+6.3)	58.3 (+33.3)	52.1 (+27.1)	47.2 (+21.5)	77.1 (-6.3)	22.9 (+6.3)	50.0 (+0.0)
EvoSkill	45.8 (+20.8)	45.8 (+20.8)	39.6 (+12.5)	41.7 (+14.6)	47.9 (+22.9)	52.1 (+27.1)	45.5 (+19.8)	83.3 (+0.0)	20.8 (+4.2)	52.1 (+2.1)
Direct Endpoint	54.2 (+29.2)	45.8 (+20.8)	47.9 (+20.8)	47.9 (+20.8)	52.1 (+27.1)	79.2 (+54.2)	54.5 (+28.8)	87.5 (+4.2)	20.8 (+4.2)	54.2 (+4.2)
WikiSkill	52.1 (+27.1)	43.8 (+18.8)	41.7 (+14.6)	52.1 (+25.0)	41.7 (+16.7)	52.1 (+27.1)	47.2 (+21.5)	85.4 (+2.1)	18.8 (+2.1)	52.1 (+2.1)
HEIRS (Ours)	**60.4** (+35.4)	45.8 (+20.8)	**60.4** (+33.3)	47.9 (+20.8)	**60.4** (+35.4)	79.2 (+54.2)	**59.0** (+33.3)	87.5 (+4.2)	20.8 (+4.2)	**54.2** (+4.2)
Full History	50.0 (+25.0)	45.8 (+20.8)	60.4 (+33.3)	56.3 (+29.2)	52.1 (+27.1)	79.2 (+54.2)	57.3 (+31.6)	83.3 (+0.0)	18.8 (+2.1)	51.0 (+1.0)
Scratch SkillOpt	45.8 (+20.8)		64.6 (+37.5)		52.1 (+27.1)		54.2 (+28.5)	87.5 (+4.2)	20.8 (+4.2)	54.2 (+4.2)

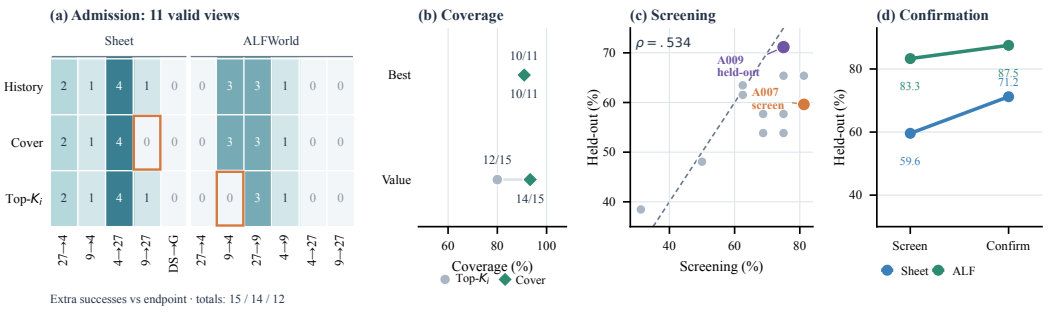

Figure 6: **Admission and selection lose historical value at different stages.** (a) Additional successes over the endpoint on the 16-task screen retained by Full History, History-Cover, and matched Source Top-K_i across 11 Spreadsheet/ALFWorld views; totals are 15, 14, and 12. (b) Inclusion of the panel-best version and retained option value after admission. (c) DeepSeek→GLM rankings on the 16-task screen and a separate 52-task characterization panel; A007 leads on the screen, whereas A009 leads on the larger panel ($\rho = .534$). (d) Confirmation changes both cross-family choices.

History-Cover retains most available value. At matched candidate budgets across the 11 completed Spreadsheet and ALFWorld views, History-Cover retains 14 of the 15 additional panel successes available in the full archive, compared with 12 for source-score Top-K_i. Most historical advantage therefore remains available after admission; realizing it still depends on reliable target-side selection. Appendix E.1 provides the complete admission analysis.

Retention does not guarantee correct selection. On DeepSeek-to-GLM Spreadsheet, A009 survives admission but the 16-task screen ranks A007 higher; a separate 52-task panel reverses their order, with only moderate rank agreement ($\rho = .534$; Figure 6c). Thus a useful retained Skill can still be mis-ranked, motivating HEIRS's disjoint confirmation stage (Figure 6d; Appendix E.2).

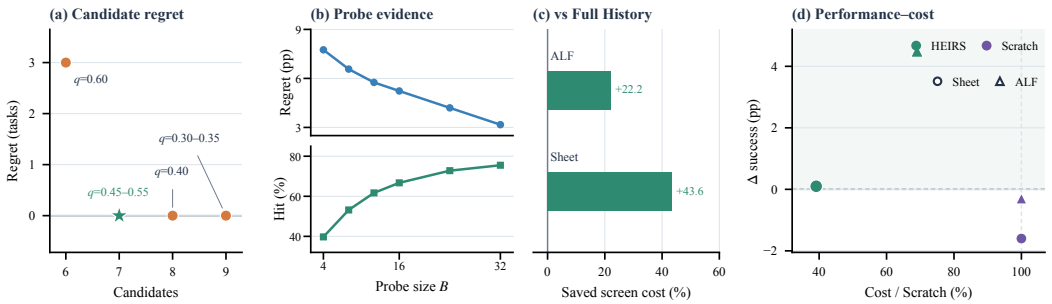

Figure 7: **Candidate coverage and evidence determine the budget–return tradeoff.** (a) Candidate count and admission regret versus q; (b) selection regret and top-set hit rate versus probe size; (c) screening savings versus Full History; (d) held-out gain versus Scratch-normalized total cost.

5.4 Target-Evaluation Budget and Cost

Figure 7 shows the tradeoff between candidate breadth and evidence per candidate. Increasing q from .45 to .60 removes A012 and loses $3/16$ successes on 4B→27B, while increasing the DeepSeek-to-GLM probe from 4 to 32 tasks reduces mean selection regret from 7.75 to 3.17 points (Appendix E.2).

Matched screening. History-Cover saves 43.6% of Spreadsheet and 22.2% of ALFWorld screening calls (704/1,248 and 448/576); confirmation adds 376 and 310 calls.

Benchmark	Direct Endpoint	HEIRS	Scratch SkillOpt
Spreadsheet	33.6 / 0	33.7 / 1080	32.0 / 2760
ALFWorld	54.5 / 0	59.0 / 758	54.2 / 1098

Table 2: **End-to-end cost.** Held-out success (%) / target evaluations; HEIRS includes screening and confirmation.

End-to-end migration cost. Including confirmation, HEIRS uses 1,838 rather than 3,858 evaluations (52.4% fewer). Mean success also rises from 32.0% to 33.7% on Spreadsheet and from 54.2% to 59.0% on ALFWorld.

6 Conclusions, Limitations, and Broader Impact

Executor replacement can change which historical Skill works best. HEIRS therefore re-evaluates the archive and transfers one Skill unchanged. Across our experiments, it improves mean Qwen ALFWorld and DeepSeek-to-GLM Spreadsheet performance while using 52.4% fewer target-side evaluations than fresh SkillOpt. The gains remain migration-specific rather than universal.

The experiments also clarify where historical revaluation succeeds or fails. History-Cover preserves most of the value available in the archive, while disjoint confirmation reduces dependence on a single small screening set. When the archive already contains a target-compatible Skill, this allows previous optimization work to be reused instead of repeated. When no such Skill exists, revaluation cannot replace target-side optimization. The practical role of HEIRS is therefore to test the archive first and optimize again only when the preserved alternatives are insufficient.

Limitations. HEIRS is bounded by both archive quality and target evidence. It cannot outperform the best useful Skill already present in the history, and a limited evaluation budget can still mis-rank retained candidates. Larger archives make this tradeoff harder because more alternatives must be compared with finite evidence. Our experiments cover SpreadsheetBench-Verified and ALFWorld with the reported executor pairs; future work should test longer histories, other Skill representations and model families, and stronger distribution shifts.

Broader impact. Reusing historical Skills can reduce repeated optimization and its associated inference cost, but it can also preserve brittle or unsafe behavior that was not exposed by the source evaluator. HEIRS should therefore be treated as a migration mechanism rather than a safety guarantee. Consequential deployments should track provenance and independently validate the selected Skill.

AI USE STATEMENT

Generative AI tools were used for literature retrieval and organization, research planning and feedback on experimental design, analysis and scripting assistance, drafting and restructuring portions of the manuscript, language editing, and LaTeX preparation. AI-assisted code, numerical summaries, and manuscript content were checked against the underlying experimental artifacts and predefined evaluation protocols. We did not use generative AI tools to generate synthetic datasets or to prove mathematical claims. The authors reviewed all AI-assisted work and take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

We will release the Skill artifacts, source histories, task manifests, per-task outcomes, resource records, and code needed to reproduce all tables and figures (Appendix F.2).

REFERENCES

Salaheddin Alzubi, Noah Provenzano, Jaydon Bingham, Christian A. Calvo, Weiyuan Chen, and Tu Vu. EvoSkill: Automated skill discovery for multi-agent systems. In *Proceedings of the 3rd Conference on Language Modeling (COLM '26)*, 2026. URL https://arxiv.org/abs/2603.02766.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations (ICLR '17)*, 2017.

Zisu Huang, Jingwen Xu, Yifan Yang, Ziyang Gong, Qihao Yang, Muzhao Tian, Xiaohua Wang, Changze Lv, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Xue Yang, Dongdong Chen, Xiaoqing Zheng, and Chong Luo. From raw experience to skill consumption: A systematic study of model-generated agent skills, 2026. URL https://arxiv.org/abs/2605.23899.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning (ICML '21)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Shixin Li, Chaoxiang He, Xiaojing Ma, Bin Benjamin Zhu, Shuo Wang, Hongsheng Hu, Dongmei Zhang, and Linchen Yu. Enhancing adversarial transferability with checkpoints of a single model's training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '25)*, pp. 20685–20694, 2025a.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning (ICML '25)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35581–35602. PMLR, 2025b. URL https://proceedings.mlr.press/v267/li25bp.html.

Yuxuan Liu, Zhaochen Su, Yuhao Zhang, Jiahe Guo, Zhongwei Xie, Huihao Jing, Lingyun Xie, Qing Zong, Yauwai Yim, Zhixiong Zhang, Haoran Li, and Yangqiu Song. Rethinking self-evolving agent skills: Feedback dynamics over multiple rounds, 2026. URL https://arxiv.org/abs/2608.02636.

Zeyao Ma, Bohan Zhang, Jing Zhang, Jifan Yu, Xiaokang Zhang, Xiaohan Zhang, Sijia Luo, Xi Wang, and Jie Tang. SpreadsheetBench: Towards challenging real world spreadsheet manipulation. In *Advances in Neural Information Processing Systems (NeurIPS '24)*, volume 37, pp. 94871–94908, 2024. doi: 10.52202/079017-3007.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Trace2Skill: Distill trajectory-local lessons into transferable agent skills, 2026. URL https://arxiv.org/abs/2603.25158.

Chengshuai Shi, Kun Yang, Zihan Chen, Jundong Li, Jing Yang, and Cong Shen. Efficient prompt optimization through the lens of best arm identification. In *Advances in Neural Information Processing Systems (NeurIPS '24)*, volume 37, pp. 99646–99685, 2024. doi: 10.52202/079017-3161. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b46bc1449205888e1883f692aff1a252-Abstract-Conference.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS '23)*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR '21)*, 2021.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL '22)*, pp. 3949–3969. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.290. URL https://aclanthology.org/2022.naacl-main.290/.

Liyan Tang, Cyrus Rashtchian, Chun-Sung Ferng, Andrew Tomkins, Da-Cheng Juan, and Tu Vu. WikiSkill: Compiling agent experience into persistent knowledge for skill evolution, 2026. URL https://arxiv.org/abs/2608.27454.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL '24), Volume 1: Long Papers*, pp. 1576–1601, 2024. doi: 10.18653/v1/2024.eacl-long.95.

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.

Jingxing Wang, Chenyu Zhou, Zhihui Fu, Jun Wang, Weiwen Liu, Weinan Zhang, and Jianghao Lin. Skills on the fly: Test-time adaptive skill synthesis for LLM agents, 2026a. URL https://arxiv.org/abs/2605.16986.

Yaxuan Wang, Quan Liu, Zhenting Wang, Zichao Li, Wei Wei, Yang Liu, and Yujia Bao. PromptBridge: Cross-model prompt transfer for large language models. In *Proceedings of the 3rd Conference on Language Modeling (COLM '26)*, 2026b. URL https://arxiv.org/abs/2512.01420.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning (ICML '25)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63897–63911. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25bx.html.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning (ICML '22)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

R. D. Wright and T. E. Ramsay, Jr. On the effectiveness of common random numbers. *Management Science*, 25(7):649–656, 1979. doi: 10.1287/mnsc.25.7.649.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations (ICLR '24)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3339f19c5fcee3ad74502947a32be9e6-Abstract-Conference.html.

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, 2026. URL https://arxiv.org/abs/2605.23904.

Jianxiang Yu, Jiapeng Zhu, Bochen Lin, Qier Cui, Zichen Ding, and Xiang Li. Skill is not one-size-fits-all: Model-aware skill alignment for LLM agents, 2026. URL https://arxiv.org/abs/2605.30723.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent architecture search via agentic supernet. In *Proceedings of the 42nd International Conference on Machine Learning (ICML '25)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 75834–75852. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25bi.html.

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 (AAAI '24)*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936.

GUIDE TO THE APPENDICES

The appendices follow the main paper's argument, from historical Skill value to selection, transfer performance, and evaluation cost. The guide below points readers to the additional analyses behind each main-paper section.

Main-paper discussion	Supplementary material
Section 3: history and mechanism	Appendix A: rankings, outcome accounting, traces, and executor dependence
Section 4: HEIRS	Appendix B: implementation, loss decomposition, and selection bounds
Section 5.1: experimental setup	Appendix C: protocols, splits, table notes, and baseline ports
Section 5.2: transfer performance	Appendix D: target-side adaptation baselines, paired uncertainty, robustness, and cross-family details
Sections 5.3–5.4: selection and budget	Appendix E: admission, ranking, and evidence-budget diagnostics
Section 5.4: migration cost	Appendix F: detailed cost accounting and reproducibility materials

A HISTORICAL REORDERING AND EXECUTION MECHANISMS

This appendix expands Section 3. It first asks whether simple source-side signals can predict which Skill will transfer well, then provides detailed task-level and execution analyses. The final subsections formalize executor dependence and present a stylized model for intuition.

A.1 CAN SOURCE-SIDE SIGNALS PREDICT THE BEST TARGET SKILL?

Checkpoint position. Figure 2 shows that the source endpoint can fall well below rank 1 after migration, while target evaluation can favor Skills that were rejected during source evolution. Checkpoint position is also inconsistent across targets: later Skills are not systematically better or worse. History position therefore provides no reliable rule for choosing the target Skill.

Skill text and edit type. We next test whether the source-side text itself provides a stable ranking signal. Two annotators score 13 anonymized Skills by the density of conditional, task-general constraints. One annotator assigns the same score to every Skill, while the other assigns nonzero scores to only two, so the resulting profiles are not informative.

We also label 154 anonymized parent–child edits as localized repairs, broad behavioral policies, or other changes. Agreement is low ($\kappa = 0.213$, below the prespecified 0.60 threshold), indicating that this edit taxonomy is not reproducible enough to support target-Skill selection.

Source-task overlap. Finally, we test whether the tasks improved on the source executor are the same tasks the target executor still needs to improve. Using A000 as a common reference on the same 16 tasks, each historical Skill S defines source-side repair and harm sets $R_s(S)$ and $H_s(S)$. From the target behavior of A000 alone, we define remaining failures D_t and preserved successes P_t. The score

$$F(S,t) = \frac{|R_s(S) \cap D_t| - |H_s(S) \cap P_t|}{16}$$

is high when source-side repairs fall on tasks the target still fails and source-side harms avoid tasks the target already solves.

This overlap score does not provide a stable target ranking. Kendall $\tau_b(F, \Delta_t)$ is -0.263 for 9B and $+0.418$ for 27B, with mean 0.078 versus the prespecified 0.30 threshold; a 10,000-draw permutation test gives one-sided $p = 0.098$. A001 and A012 both have $F = 0$ for target 27B even though the complete 280-task evaluation separates them by 26 successes. Simple source-task overlap therefore does not explain which historical Skill will be best after migration.

Table 3: **Source-side predictors of target ranking.** We test several signals available before target evaluation. None provides a reliable rule for ranking historical Skills across executors.

Hypothesis	Prespecified diagnostic	Result
Checkpoint position / provenance	Common-history source/target views	Evolution index changes sign across migrated views; source endpoint can move far below rank 1.
Skill-text profile	Blind two-annotator 13-Skill profile	One score vector is constant; the other is nonzero for only two Skills, so the ranking signal is uninformative.
Edit type	Blind labeling of 154 parent–child edits	Cohen's $\kappa = 0.213$, below the prespecified 0.60 agreement threshold.
Source-task overlap	Common 16-task source/target comparison	$\tau_{9B} = -0.263$, $\tau_{27B} = 0.418$, mean 0.078; 10k permutation $p = 0.098$.

Taken together, checkpoint position, text features, edit type, and source-task overlap do not provide a reliable source-only ranking rule. Target-side evidence is therefore needed to decide which historical Skill works best with the replacement executor.

Table 4: Six directed Qwen3.5 history–target views on within-view common 16-task diagnostic panels. "Finite-set best" is the best observed historical state on that panel.

Source	Target	Endpoint	Finite-set best	Gap
4B	9B	4/16	6/16	+12.50 pp
4B	27B	6/16	11/16	+31.25 pp
9B	4B	2/16	4/16	+12.50 pp
9B	27B	10/16	13/16	+18.75 pp
27B	4B	2/16	4/16	+12.50 pp
27B	9B	6/16	7/16	+6.25 pp

Some rows share the same source history, so the six views should be read as descriptive comparisons rather than independent replications. Each row reports one 16-task panel for the stated source–target direction.

Do source-side scores track target rankings? On the 4B history, the Spearman correlations between Skill rankings under 4B, 9B, and 27B are small: 0.035 for 4B–9B, 0.153 for 4B–27B, and -0.034 for 9B–27B. Ranking stability within each 16-task panel is also weak under split-half evaluation, so these values should be read as evidence of reordering rather than precise estimates of a population correlation. Source validation score correlates only -0.013 with the 9B ranking and -0.067 with the 27B ranking, while checkpoint index correlates -0.416 and 0.121. These source-side signals therefore provide little guidance about the target ordering.

Source-validation granularity and equal scores. Source Top-K_i uses recorded source validation as its admission signal. Table 5 gives completed Spreadsheet and ALFWorld histories: with 18–40 validation tasks, score resolution is coarse and equal scores are common. In Spreadsheet-4B, two equally scored states compete for one remaining Top-7 slot, so validation alone cannot distinguish them.

Table 5: Source-validation granularity in representative histories. Resolution is one validation task in percentage points. "Cutoff ambiguity" describes equal scores affecting matched Top-K_i admission.

Benchmark	Source	n_{val}	Resolution	Largest equal-score group	Cutoff ambiguity
Spreadsheet	4B	40	2.50	5 states at 22.5%	Top-7 has 2 equal-score states for 1 remaining slot
ALFWorld	4B	18	5.56	4 states at 50.0%	archive has only five states and all are retained

Source Top-K_i therefore measures source-ranking admission under the matched K_i budget, while HEIRS uses history coverage for admission and reserves target ranking for revaluation.

A.2 Executor Response and Outcome Accounting

The executor dependence in Section 3.2 has a population-level form. On a common task distribution, let $\mu_m(S)$ be utility under executor condition $m \in \{s, t\}$ and $e = S_{\text{end}}$. Define the source gap $a_s(S) = \mu_s(S) - \mu_s(e)$ and the relative executor response

$$c_{s\to t}(S) = \big[\mu_t(S) - \mu_s(S)\big] - \big[\mu_t(e) - \mu_s(e)\big]. \tag{10}$$

Then

$$\boxed{\mu_t(S) - \mu_t(e) = a_s(S) + c_{s\to t}(S).} \tag{11}$$

A source disadvantage reverses when the executor change favors the historical version enough to offset it. This identity describes the interaction required for population reversal; finite gate scores alone do not establish it. Appendix A.5 separates executor response from source-selection error and specifies the executor condition, including its runtime.

For the recovery–interference decomposition, let Y_m^0 and Y_m^S denote binary NoSkill and Skill outcomes on a common task distribution. Then $p_m = \Pr(Y_m^0 = 1)$, $r_m(S) = \Pr(Y_m^S = 1 \mid Y_m^0 = 0)$, and $h_m(S) = \Pr(Y_m^S = 0 \mid Y_m^0 = 1)$. Equation 3 follows by conditioning on the NoSkill outcome. For two versions under the same executor,

$$\mu_m(S) - \mu_m(S') = (1 - p_m)\big[r_m(S) - r_m(S')\big] - p_m\big[h_m(S) - h_m(S')\big]. \tag{12}$$

On the 280-task 27B panel, A001 recovers 37 NoSkill failures and disrupts 32 native successes, solving 125 tasks. A012 recovers 48 and disrupts 17, solving 151. Both use 120 NoSkill successes.

The detailed comparisons support the same conclusion at different scales. Among 27 source tasks that both A001 and A005 fail, A005 improves five tasks at the cell level and regresses on two. A012 exceeds A001 by $2/64$ on a fresh 4B set and by $3/16$, $5/48$, and $26/280$ under 27B. Because these task sets differ, they show the existence of useful historical alternatives but are not used to estimate a causal executor interaction.

A.3 Executor-Conditioned Mechanism

Matched source–target interaction. We compare selected historical Skills on the same 128 SpreadsheetBench held-out tasks under source 4B and each target executor. Using the same tasks removes differences caused simply by task composition, allowing the interaction with the executor to be measured directly. These analyses are separate from HEIRS selection. Table 6 reports paired task-bootstrap estimates.

Table 6: Matched held-out source–target interaction control on the same 128 tasks. Effects are percentage-point differences between the historical alternative and the source endpoint; the interaction subtracts the source-executor effect from the target-executor effect.

Contrast	Source 4B	Target	Interaction	Paired bootstrap 95% CI
A003−A001, target 9B	−7.03	+2.34	+9.38	[0.00, 19.53]
A012−A001, target 27B	0.00	+5.47	+5.47	[−6.25, 17.19]

The 9B interaction interval touches zero and the 27B interval crosses zero, so these controls do not establish a consistent population-level reversal. They do show that the relative effect of a historical Skill can change with the executor.

Task-level recovery and interference. For mechanism characterization, we evaluate the 4B source history on the complete 280-task SpreadsheetBench-Verified test set under the 27B executor and reuse the 200-task expanded held-out matrices for two additional directions. These task sets are separate from candidate admission, Stage-1 screening, and Stage-2 confirmation, so the analysis characterizes task-level behavior rather than influencing the inherited Skill.

Table 7: **Recovery–interference accounting on larger task sets.** NoSkill is the target executor's success count on the same tasks. Recovery and interference are defined relative to those paired NoSkill outcomes. Prefixes distinguish equally numbered artifacts from different source histories.

Setting	State	N	NoSkill	Skill	Rec./Int.
4B history → 27B, full test	4B:A001	280	120	125	37 / 32
	4B:A012	280	120	**151**	**48 / 17**
4B → 9B, expanded held-out	4B:A001	200	57	63	25 / 19
	4B:A012	200	57	51	18 / 24
27B → 9B, expanded held-out	27B:A001	200	57	60	27 / 24
	27B:A002	200	57	61	25 / 21

The full-test A012–A001 difference under target 27B is $+26$ tasks, exactly $+11$ additional recoveries plus 15 fewer interferences. The expanded 4B→9B direction provides the converse case: A012 is 12 tasks below A001 because it has seven fewer recoveries and five more interferences. In 27B→9B, A002 exceeds its source endpoint by one task despite two fewer recoveries because it introduces three fewer interferences. Both recovery and interference are therefore needed to explain migration gains.

The same fixed 4B:A001 Skill also has different task-level effects under 9B and 27B on the common 200-task matrix. Its signed effect correlation is 0.098; recovery-set Jaccard is 0.143 (7 shared tasks) and interference-set Jaccard is 0.105 (4 shared tasks). Of the 200 tasks, 81 are changed by A001 under at least one executor and six have opposite-sign effects. These outcomes illustrate Eq. 12: executor replacement changes both native capability and the tasks a fixed Skill helps or disrupts.

The larger 280-task matrix shows that A012's 27B advantage comes from both more recovered failures and fewer disrupted successes. This larger sample provides a more stable mechanism picture than the smaller 48-task view.

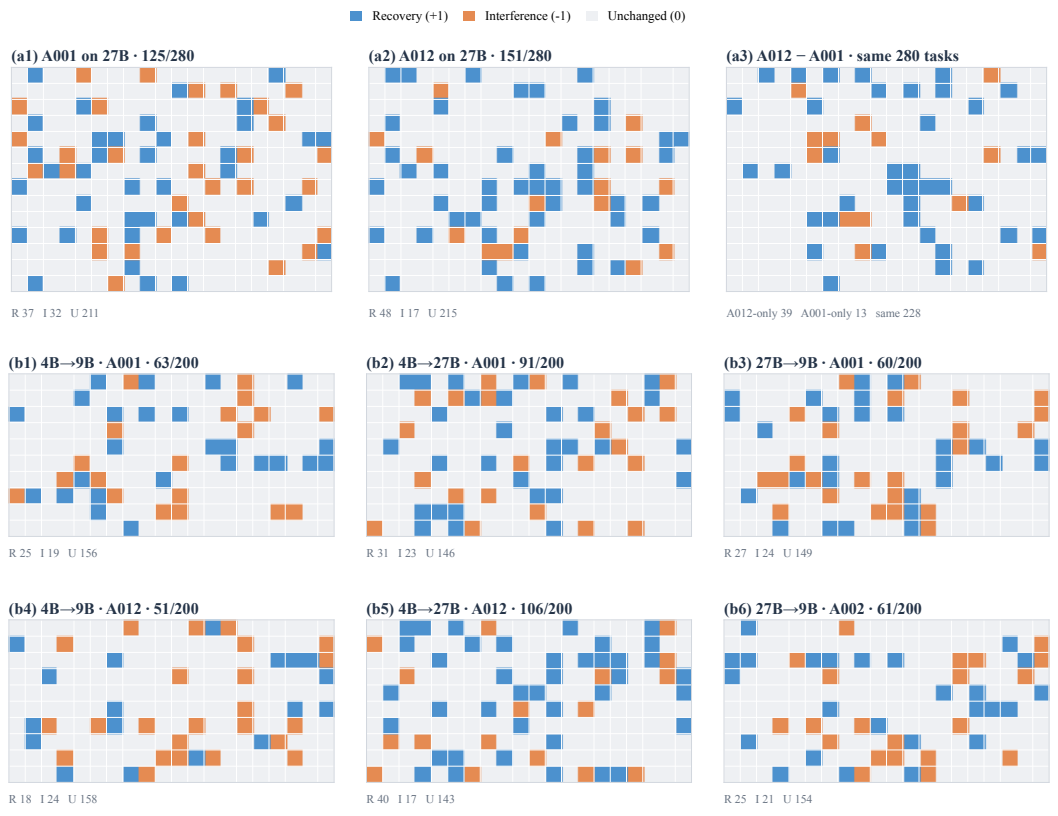

Figure 8: **Expanded task-level intervention maps.** Panels (a1–a3) compare A001, A012, and NoSkill on the same 280 target-27B tasks; (b1–b6) show three 200-task migrations with shared ordering. Blue indicates recovery, orange interference, and gray unchanged outcomes.

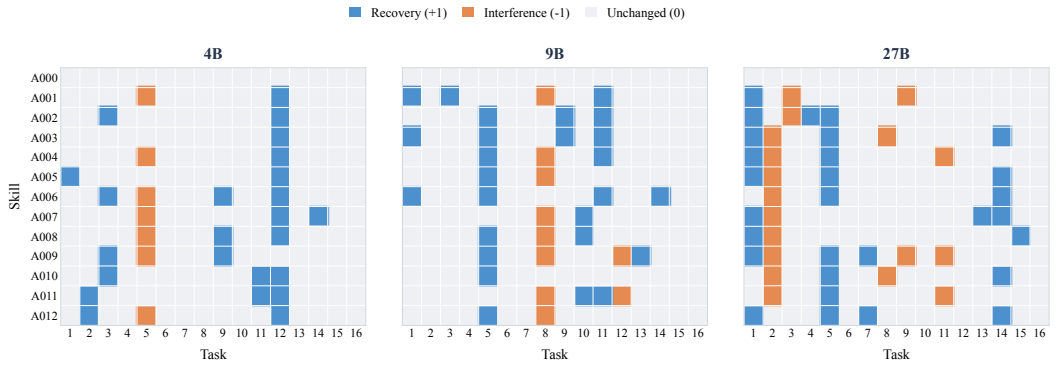

Figure 9: **Whole-history task effects on a shared 16-task panel.** Rows are archived Skills, columns are task identities, and each cell is the hard-outcome change from A000. The same history produces different effect patterns under 4B, 9B, and 27B.

A.4 Correct Computation and Output Delivery

Panel and artifact coverage. We compare A001 and A012 on the same 128 SpreadsheetBench tasks under 9B and 27B. Complete output artifacts are available for 124 tasks; the four remaining tasks fail under all four conditions and are kept in the success counts but omitted from artifact-level analysis. Because the executor condition includes the serving and execution environment, this comparison should not be interpreted as an isolated effect of model size.

A directly observable failure signature. We mark a condition only when its final verification report identifies a first wrong value as `None` and the corresponding cell in the saved output workbook contains a formula element with an empty cached value. This checks the actual workbook XML rather than a keyword in generated code. The marker is not an exhaustive formula-error count: earlier failures may mask it, and the formula itself may be wrong.

Some marked cases also contain incorrect formula logic, so the disappearance of an uncached formula does not imply that recalculation alone would fix every task.

Contrary cases and scope. We inspect a stratified set of traces covering opposite A001–A012 changes across executors, joint improvements and regressions, stable successes, and stable failures. These cases help distinguish the observed output-path failure from broader differences in task understanding.

The contrary cases do not support a simple explanation that the larger executor always applies instructions more carefully. Different tasks fail for different reasons: wrong cell indexing, writing references instead of values, overly broad range edits, output-order changes, and unresolved table-structure errors. The output-path mechanism is therefore one concrete failure mode, not a general theory of executor capability.

A one-assignment intervention. For task 57117, the original 27B A001 program computes `result` from substring tests for BA and MA; for task 51431, it computes `first_name` by splitting at the first semicolon. Both write formulas instead of these values. We change one assignment:

Task	Original assignment	Replacement
57117	`target_cell.value = formula`	`target_cell.value = result`
51431	`target_cell.value = formula`	`target_cell.value = first_name`

The original A001 programs fail, while the one-assignment modifications pass all required cells; the corresponding A012 programs also pass. The inputs and evaluator are unchanged, and no new model generation is involved. For these two fixed programs, output assignment is therefore sufficient to explain the failure. This intervention does not identify which natural-language Skill instruction caused the model to choose that output path.

Interpretation. This result separates computing the right answer from delivering that answer in the form expected by the evaluator. It does not show that one specific Skill sentence caused the behavior, nor does it imply that every formula-related failure has the same cause. A direct instruction-level ablation would be required to make that stronger claim.

A.5 Executor-Relative Historical Value

Fix the observed archive $\mathcal{H}_s$, its source-selected endpoint $e = S_{\text{end}}$, and a task distribution P. For executor condition $m \in \{s, t\}$, define

$$\mu_m(S) = \mathbb{E}_{x \sim P, \xi_m}[R(S, M_m, x; \xi_m)] \in [0, 1], \tag{13}$$

where ξ_m represents execution randomness under the recorded configuration. The same task distribution is used for source–target comparison. The executor condition includes the execution interface, so $c_{s \to t}$ also reflects runtime and harness differences.

Let $a_s(S) = \mu_s(S) - \mu_s(e)$ and

$$c_{s \to t}(S) = \big[\mu_t(S) - \mu_s(S)\big] - \big[\mu_t(e) - \mu_s(e)\big]. \tag{14}$$

Equivalently, if $g_s(S) = \mu_s(e) - \mu_s(S)$ and $d(S) = \mu_t(S) - \mu_s(S)$, then $a_s(S) = -g_s(S)$ and $c_{s \to t}(S) = d(S) - d(e)$.

Proposition 1 (Historical value under executor change)**.** *For every $S \in \mathcal{H}_s$,*

$$\mu_t(S) - \mu_t(e) = a_s(S) + c_{s \to t}(S) = d(S) - d(e) - g_s(S). \tag{15}$$

Consequently, a state with nonnegative source gap $g_s(S) \geq 0$ outperforms the endpoint under the target condition exactly when $c_{s \to t}(S) > g_s(S)$.

Proof. Add and subtract $\mu_s(S)$ and $\mu_s(e)$ in $\mu_t(S) - \mu_t(e)$. Regrouping gives both claims. □

The identity distinguishes relative executor response from a uniform change in model quality. If $c_{s\to t}(S) = 0$ for every state, population differences relative to e are preserved. The endpoint is selected by finite source validation and need not maximize μ_s. Positive target option value can therefore also reflect source-selection error; a population reversal requires separate evidence.

A.6 Stylized Source–Target Alignment Model

Proposition 2 (Source-optimality under imperfect executor alignment)**.** *Let $\{(X_j, Y_j)\}_{j=1}^{L}$ be i.i.d. standardized jointly Gaussian source and target population utilities with $\mathrm{Corr}(X_j, Y_j) = \rho$. Define $I_s \in \arg\max_j X_j$, $I_t \in \arg\max_j Y_j$, and $a_L = \mathbb{E}[\max_{j\le L} Z_j]$ for i.i.d. $Z_j \sim \mathcal{N}(0,1)$. Then*

$$\mathbb{E}[Y_{I_s}] = \rho a_L, \qquad \mathbb{E}[Y_{I_t} - Y_{I_s}] = (1-\rho)a_L. \tag{16}$$

If target utility is $\mu_t^0 + \sigma_t Y_j$, the expected target-history regret is $\sigma_t(1-\rho)a_L$.

Corollary 1 (Finite source-validation noise)**.** *Under Proposition 2, let the source selector observe $\widetilde{X}_j = X_j + \varepsilon_j$, where $\varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2)$ independently, and return $\widetilde{I}_s \in \arg\max_j \widetilde{X}_j$. Then*

$$\mathbb{E}[Y_{\widetilde{I}_s}] = \frac{\rho}{\sqrt{1+\tau^2}} a_L, \qquad \mathbb{E}[Y_{I_t} - Y_{\widetilde{I}_s}] = \left(1 - \frac{\rho}{\sqrt{1+\tau^2}}\right) a_L. \tag{17}$$

The model isolates source-selection effects under an i.i.d. Gaussian utility landscape; the observed Skill trajectory falls outside this simplified model.

Proof of Proposition 2. For $|\rho| < 1$, a standardized jointly Gaussian pair can be written as

$$Y_j = \rho X_j + \sqrt{1-\rho^2}\,\varepsilon_j, \tag{18}$$

where $\varepsilon_1, \dots, \varepsilon_L$ are i.i.d. standard Gaussians independent of $X_1, \dots, X_L$. Because $I_s \in \arg\max_j X_j$ is measurable with respect to the source utilities,

$$\mathbb{E}[Y_{I_s} \mid X_1, \dots, X_L] = \rho X_{I_s} + \sqrt{1-\rho^2}\,\mathbb{E}[\varepsilon_{I_s} \mid X_1, \dots, X_L] \tag{19}$$

$$= \rho X_{I_s}. \tag{20}$$

Taking expectation yields

$$\mathbb{E}[Y_{I_s}] = \rho\,\mathbb{E}[\max_j X_j] = \rho a_L. \tag{21}$$

The target utilities $Y_1, \dots, Y_L$ are themselves i.i.d. standard Gaussian, hence

$$\mathbb{E}[Y_{I_t}] = \mathbb{E}[\max_j Y_j] = a_L. \tag{22}$$

Linearity of expectation therefore gives

$$\mathbb{E}[Y_{I_t} - Y_{I_s}] = (1-\rho)a_L. \tag{23}$$

The degenerate cases $\rho = \pm 1$ satisfy the same expression directly. If target utilities are $U_j^t = \mu_t^0 + \sigma_t Y_j$, the additive mean cancels and the regret is multiplied by σ_t. Finally, as $L \to \infty$, the standard Gaussian maximum obeys $a_L/\sqrt{2\log L} \to 1$, giving the first-order scaling of the stylized regret. □

Proof of Corollary 1. Let $Z_j = \widetilde{X}_j/\sqrt{1+\tau^2}$. Then $Z_j \sim \mathcal{N}(0,1)$, the ordering of Z_j is identical to that of $\widetilde{X}_j$, and

$$\mathrm{Corr}(Z_j, Y_j) = \frac{\mathrm{Cov}(X_j + \varepsilon_j, Y_j)}{\sqrt{1+\tau^2}} = \frac{\rho}{\sqrt{1+\tau^2}}. \tag{24}$$

Applying Proposition 2 to the jointly Gaussian pair (Z_j, Y_j) gives

$$\mathbb{E}[Y_{\widetilde{I}_s}] = \frac{\rho}{\sqrt{1+\tau^2}} a_L, \tag{25}$$

and subtracting from $\mathbb{E}[Y_{I_t}] = a_L$ yields Eq. 17. □

The proposition and corollary separate two effects in the stylized model: source and target utilities may be imperfectly aligned, and finite source validation may select the wrong endpoint. Real Skill trajectories are dependent and nonstationary, so we use this model only for intuition rather than as a fitted description of the experiments. The observed B16 correlations are therefore descriptive diagnostics; ρ and τ remain population-level stylized parameters.

B HEIRS Implementation and Selection Analysis

This appendix provides the formal details behind Section 4. We first define where historical value can be lost, then give the HEIRS implementation and analyze candidate admission and target-side ranking. The probability bounds apply to fresh i.i.d. target tasks; the deterministic Task-Cover rule used in HEIRS is evaluated empirically.

B.1 Admission, Screening, and Final-Ranking Losses

A policy admits $S_{\text{end}} \in \mathcal{C} \subseteq \mathcal{H}_s$, screens to $\mathcal{C}_{\text{short}}$, and selects $\widehat{S} \in \mathcal{C}_{\text{short}}$. The losses on D are

$$R_{\text{cand}}(D) = \max_{S \in \mathcal{H}_s} U(S; M_t, D) - \max_{S \in \mathcal{C}} U(S; M_t, D), \tag{26}$$

$$R_{\text{screen}}(D) = \max_{S \in \mathcal{C}} U(S; M_t, D) - \max_{S \in \mathcal{C}_{\text{short}}} U(S; M_t, D), \tag{27}$$

$$R_{\text{conf}}(D) = \max_{S \in \mathcal{C}_{\text{short}}} U(S; M_t, D) - U(\widehat{S}; M_t, D). \tag{28}$$

Here R_{cand} is loss before target evaluation, R_{screen} is loss from removing an admitted candidate, and R_{conf} is loss from the final choice among survivors. Their sum gives Equation 4. These are evaluation quantities; held-out scores are never supplied to the selector.

B.2 Recoverable Value and Endpoint Anchoring

For any finite real-valued utility vector $u : \mathcal{H}_s \to \mathbb{R}$ (the empirical success utilities used in the experiments lie in $[0, 1]$), any retained set $e \in \mathcal{C} \subseteq \mathcal{H}_s$, and any selected $\widehat{S} \in \mathcal{C}$, define

$$V^u = \max_{S \in \mathcal{H}_s} u(S) - u(e), \qquad R^u_{\text{cand}} = \max_{S \in \mathcal{H}_s} u(S) - \max_{S \in \mathcal{C}} u(S), \tag{29}$$

$$V^u_{\mathcal{C}} = \max_{S \in \mathcal{C}} u(S) - u(e), \qquad R^u_{\text{sel}} = \max_{S \in \mathcal{C}} u(S) - u(\widehat{S}), \tag{30}$$

$$G^u = u(\widehat{S}) - u(e). \tag{31}$$

Proposition 3 (Exact recoverable-value decomposition)**.** *For every finite utility vector u,*

$$V^u_{\mathcal{C}} = V^u - R^u_{\text{cand}} \geq 0, \qquad G^u = V^u_{\mathcal{C}} - R^u_{\text{sel}} = V^u - R^u_{\text{cand}} - R^u_{\text{sel}}. \tag{32}$$

In particular, if selection is exact within the retained set ($R^u_{\text{sel}} = 0$), endpoint-anchored inheritance cannot underperform the endpoint. Conversely, $G^u < 0$ if and only if $R^u_{\text{sel}} > V^u_{\mathcal{C}}$.

Proof. Add and subtract $\max_{S \in \mathcal{C}} u(S)$. Since $e \in \mathcal{C}$, $\max_{S \in \mathcal{C}} u(S) \geq u(e)$, hence $V^u_{\mathcal{C}} \geq 0$. The remaining statements are immediate rearrangements. □

The proposition gives a direct interpretation of the two failure modes. When $R_{\text{cand}} > 0$, shortlisting has removed some available historical value before target evaluation. The remaining loss comes from choosing the wrong retained version. The endpoint remains a candidate. A held-out result below Direct Endpoint therefore means that selection error exceeds the retained advantage.

Corollary 2 (Stylized end-to-end recovery)**.** *Under Corollary 1, let $e = \widetilde{I}_s$ and target utility be $\mu^0_t + \sigma_t Y_j$. Any retained set containing e and any historical selector satisfy*

$$\mathbb{E}\Big[\mu_t(\widehat{S}) - \mu_t(\widetilde{I}_s)\Big] = \sigma_t \left(1 - \frac{\rho}{\sqrt{1+\tau^2}}\right) a_L - \mathbb{E}\big[R^t_{\text{cand}} + R^t_{\text{sel}}\big]. \tag{33}$$

Proof of Corollary 2. Under Corollary 1 with $e = \widetilde{I}_s$ and target utility $u_j = \mu_t^0 + \sigma_t Y_j$,

$$\mathbb{E}[V^u] = \sigma_t \, \mathbb{E}[Y_{I_t} - Y_{\widetilde{I}_s}] = \sigma_t \left(1 - \frac{\rho}{\sqrt{1+\tau^2}}\right) a_L. \tag{34}$$

The exact identity in Proposition 3 holds pointwise for every realized utility vector and selected state. Taking expectations therefore gives

$$\mathbb{E}[G^u] = \sigma_t \left(1 - \frac{\rho}{\sqrt{1+\tau^2}}\right) a_L - \mathbb{E}[R^u_{\text{cand}} + R^u_{\text{sel}}]. \tag{35}$$

□

B.3 HEIRS Implementation

Candidate admission. Each byte-distinct Skill is represented by concatenated normalized word TF–IDF $(1, 2)$ and character `char_wb` TF–IDF $(3, 5)$ features. History-Cover applies cosine distance and average-linkage clustering, always retains the source endpoint, and keeps the medoid and farthest boundary member of each cluster. We test within-history distance quantiles $\{.30, .35, .40, .45, .50, .55, .60\}$ on development data and use $q = .45$, the coarsest tested setting with zero observed candidate regret.

Stage 1: screening. Task-Cover constructs one 16-task panel from instruction text without using candidate outcomes. Every admitted Skill is evaluated on the same tasks. HEIRS keeps all Skills from the two highest distinct hard-score levels, including ties, and caps the shortlist at four. If more than four Skills qualify, Stage-1 score, source validation, and chronology resolve the overflow. Screening scores are used only to form the shortlist.

Stage 2: confirmation. The shortlist is evaluated on a benchmark-native target set disjoint from Stage 1 and final held-out evaluation. All shortlisted Skills receive the same first 16 confirmation tasks. If the leaders tie, confirmation continues with two eight-task blocks on Spreadsheet (up to 32 tasks total) or the final two validation tasks on ALFWorld. At the budget cap, remaining ties prefer the source endpoint when present, then the Stage-1 leader, source validation, and chronology. Skills removed after Stage 1 are not reconsidered. The choices $q = .45$ and $B = 16$ are empirical operating points; confirmation reduces dependence on the screening set but does not guarantee population-optimal selection.

B.4 Limit of Source-Only Admission

A source-only admission rule can reduce target cost, but without any assumption connecting source information to target utility, it cannot guarantee that the best target Skill is retained. The following proposition states this limitation formally.

Proposition 4 (No-free-lunch for strict source-only admission)**.** *Fix any proper retained set $e \in \mathcal{C} \subsetneq \mathcal{H}_s$ constructed without target-utility information. Without an assumption linking the information used for admission to target utility, there exists a utility vector $u : \mathcal{H}_s \to [0, 1]$ for which*

$$V^u = R^u_{\text{cand}} = 1, \qquad V^u_{\mathcal{C}} = 0. \tag{36}$$

Therefore no strict source-only admission rule can guarantee nonzero retained option value, or zero candidate regret, uniformly over arbitrary target utilities.

Proof. Choose any omitted state $S_o \in \mathcal{H}_s \setminus \mathcal{C}$. Set $u(S_o) = 1$ and $u(S) = 0$ for every other state. Because $e \in \mathcal{C}$, the archive optimum is one and every retained state has utility zero. □

History-Cover must therefore be judged empirically: we measure its candidate loss R_{cand} and compare it with alternative rules at the same candidate count. A stronger guarantee would require an explicit assumption linking source-side information to target performance.

Nested candidate sets. For nested sets $\mathcal{C}_1 \subseteq \mathcal{C}_2$, maximizing the same utility over the larger set cannot decrease the retained optimum, so $R^u_{\text{cand}}(\mathcal{C}_2) \leq R^u_{\text{cand}}(\mathcal{C}_1)$. Candidate count alone does not imply this ordering for non-nested sets. Larger sets can also make selection harder. With noisy probe scores, Full History need not outperform a smaller candidate set on held-out tasks.

B.5 Relative Probe Distortion and Selection Stability

We now analyze one ranking stage over a fixed candidate set. The result asks how much selection quality can degrade when the observed ranking score differs from the utility we ultimately care about. HEIRS applies this reasoning separately to screening and confirmation rather than treating its two stages as one argmax. Let $\widehat{\mu}(S)$ be a score used to choose $\widehat{S} \in \arg\max_{S\in\mathcal{C}} \widehat{\mu}(S)$. For a comparison utility u, define the score error

$$\epsilon_u(S) = \widehat{\mu}(S) - u(S), \qquad \Omega_{\mathcal{C}}^u = \mathrm{osc}_{\mathcal{C}}(\epsilon_u) = \max_{S\in\mathcal{C}} \epsilon_u(S) - \min_{S\in\mathcal{C}} \epsilon_u(S). \tag{37}$$

When $u = \mu_t$ we write ϵ_t and $\Omega_{\mathcal{C}}$ for brevity.

Proposition 5 (Relative-error control for a fixed ranking stage)**.** *For any finite utility vector u,*

$$R_{\mathrm{sel}}^u \le \Omega_{\mathcal{C}}^u, \qquad G^u \ge V^u - R_{\mathrm{cand}}^u - \Omega_{\mathcal{C}}^u. \tag{38}$$

Moreover, with $m_{\mathrm{probe}} = \widehat{\mu}(\widehat{S}) - \widehat{\mu}(e) \ge 0$,

$$G^u \ge m_{\mathrm{probe}} - \Omega_{\mathcal{C}}^u \ge -\Omega_{\mathcal{C}}^u. \tag{39}$$

Proof. Choose $S^\star \in \arg\max_{S\in\mathcal{C}} u(S)$. Since $\widehat{\mu}(\widehat{S}) \ge \widehat{\mu}(S^\star)$,

$$u(S^\star) - u(\widehat{S}) = \widehat{\mu}(S^\star) - \widehat{\mu}(\widehat{S}) + \epsilon_u(\widehat{S}) - \epsilon_u(S^\star) \tag{40}$$

$$\le \epsilon_u(\widehat{S}) - \epsilon_u(S^\star) \le \Omega_{\mathcal{C}}^u. \tag{41}$$

Substitute into Proposition 3. For the endpoint-relative form,

$$u(\widehat{S}) - u(e) = \widehat{\mu}(\widehat{S}) - \widehat{\mu}(e) - \big[\epsilon_u(\widehat{S}) - \epsilon_u(e)\big] \tag{42}$$

$$\ge m_{\mathrm{probe}} - \Omega_{\mathcal{C}}^u. \tag{43}$$

Because $e \in \mathcal{C}$ and $\widehat{S}$ maximizes the probe score, $m_{\mathrm{probe}} \ge 0$. □

The bound depends only on errors that change candidates relative to one another. A common additive shift applied to every candidate cancels out and cannot change the selected Skill.

Corollary 3 (Ranking stability under relative distortion)**.** *Let $u^\star = \max_{S\in\mathcal{C}} u(S)$ and define the gap from the optimal utility level to the best strictly suboptimal retained state by*

$$\gamma_{\mathcal{C}}^u = u^\star - \max_{S\in\mathcal{C}:\, u(S)<u^\star} u(S), \tag{44}$$

with $\gamma_{\mathcal{C}}^u = +\infty$ when all retained states have equal utility. If $\Omega_{\mathcal{C}}^u < \gamma_{\mathcal{C}}^u$, then every maximizer of $\widehat{\mu}$ is also a maximizer of u over $\mathcal{C}$, and $R_{\mathrm{sel}}^u = 0$.

Proof. Take any u-optimal state $S^\star$ and any strictly suboptimal state S. Then

$$\widehat{\mu}(S^\star) - \widehat{\mu}(S) = u(S^\star) - u(S) + \epsilon_u(S^\star) - \epsilon_u(S) \ge \gamma_{\mathcal{C}}^u - \Omega_{\mathcal{C}}^u > 0. \tag{45}$$

Hence no strictly suboptimal state can maximize the probe score. □

Thus the selector does not need perfectly calibrated utility estimates; it only needs to preserve the ordering of the leading candidates.

B.6 Common-Probe Selection under Random Tasks

The next result quantifies how the number of shared target tasks affects ranking error when those tasks are fresh i.i.d. draws. It is a statistical model for common-task evaluation, not a guarantee for deterministic Task-Cover or the full two-stage HEIRS procedure. Condition on fixed $\mathcal{H}_s$, $\mathcal{C}$, and e constructed independently of the new target observations. Let $Y_1, \dots, Y_B$ be independent, identically distributed task-wise reward vectors with $Y_i(S) \in [0,1]$ for every $S \in \mathcal{C}$. Dependence among candidates within the same vector Y_i is unrestricted, matching the fact that every retained Skill is evaluated on the same task identity. Write

$$\mu_Q(S) = \mathbb{E}[Y_i(S)], \qquad \widehat{\mu}_B(S) = \frac{1}{B}\sum_{i=1}^{B} Y_i(S). \tag{46}$$

Let $S_Q^\star \in \arg\max_{S\in\mathcal{C}} \mu_Q(S)$ and $\Delta_S = \mu_Q(S_Q^\star) - \mu_Q(S)$.

Proposition 6 (Margin-sensitive common-probe selection)**.** *For every strictly suboptimal candidate S with $\Delta_S > 0$,*

$$\Pr\{\widehat{S} = S\} \le \exp(-B\Delta_S^2/2). \tag{47}$$

Consequently,

$$\Pr\{\widehat{S} \notin \arg\max_{S\in\mathcal{C}} \mu_Q(S)\} \le \sum_{S:\Delta_S>0} \exp(-B\Delta_S^2/2), \tag{48}$$

$$\mathbb{E}[R_{\mathrm{sel}}^{\mu_Q}] \le \sum_{S:\Delta_S>0} \Delta_S \exp(-B\Delta_S^2/2). \tag{49}$$

Proof. For a fixed suboptimal S, define the paired difference $Z_i = Y_i(S_Q^\star) - Y_i(S) \in [-1, 1]$, with $\mathbb{E}[Z_i] = \Delta_S$. If the selector returns S, then $\widehat{\mu}_B(S) \ge \widehat{\mu}_B(S_Q^\star)$, including possible score equality for which the resolution rule selects S, so $B^{-1}\sum_i Z_i \le 0$. Hoeffding's inequality Hoeffding (1963) gives Eq. 47. A union bound gives Eq. 48. Finally, $R_{\mathrm{sel}}^{\mu_Q} = \Delta_{\widehat{S}}$. Summing each suboptimal state's contribution $\Delta_S \Pr\{\widehat{S} = S\}$ gives the expected-regret bound in Eq. 49. □

The bound depends on both retained-candidate utility gaps and the probe size B.

Probe–deployment mismatch. Let μ_P be deployment utility under the same target execution condition and define

$$\beta_{\mathcal{C}} = \mathrm{osc}_{\mathcal{C}}(\mu_Q - \mu_P). \tag{50}$$

For any selected state,

$$R_{\mathrm{sel}}^{\mu_P} \le R_{\mathrm{sel}}^{\mu_Q} + \beta_{\mathcal{C}}. \tag{51}$$

Indeed, for any $S_P^\star \in \arg\max_{S\in\mathcal{C}} \mu_P(S)$ and $h = \mu_Q - \mu_P$,

$$R_{\mathrm{sel}}^{\mu_P} = \mu_Q(S_P^\star) - \mu_Q(\widehat{S}) + h(\widehat{S}) - h(S_P^\star) \tag{52}$$

$$\le R_{\mathrm{sel}}^{\mu_Q} + \beta_{\mathcal{C}}. \tag{53}$$

More probe tasks can therefore reduce sampling noise, but they cannot fix a systematic mismatch between the probe-task distribution and deployment.

Corollary 4 (A distribution-free-in-margin high-probability form)**.** *Condition on fixed $\mathcal{C}$ and e constructed independently of the sampled target observations, let $K = |\mathcal{C}| \ge 2$, let $\delta \in (0, 1)$, and define $G^{\mu_P} = \mu_P(\widehat{S}) - \mu_P(e)$. With probability at least $1 - \delta$,*

$$G^{\mu_P} \ge V^{\mu_P} - R_{\mathrm{cand}}^{\mu_P} - \beta_{\mathcal{C}} - \sqrt{\frac{2\log((K-1)/\delta)}{B}}. \tag{54}$$

For $K = 1$, $\mathcal{C} = \{e\}$ and the selected gain is identically zero.

Proof. For any $\eta > 0$, the event $R_{\mathrm{sel}}^{\mu_Q} \ge \eta$ requires selection of some candidate with $\Delta_S \ge \eta$. Proposition 6 and a union bound over at most $K - 1$ strictly suboptimal candidates give

$$\Pr\{R_{\mathrm{sel}}^{\mu_Q} \ge \eta\} \le (K-1)\exp(-B\eta^2/2). \tag{55}$$

Set $\eta = \sqrt{2\log((K-1)/\delta)/B}$, use Eq. 51, and substitute into Proposition 3. □

Breadth–evidence tradeoff under a fixed total budget. Separately from the current HEIRS protocol, consider an evaluator with a fixed logical target-execution budget $T \ge K$ and an externally chosen candidate-set size K. It can fund $B(K) = \lfloor T/K \rfloor$ complete common-task evaluations per candidate, so each pairwise term in Eq. 48 becomes

$$\exp\left(-\frac{\lfloor T/K \rfloor \Delta_S^2}{2}\right). \tag{56}$$

Widening a nested candidate set can reduce R_{cand} while adding near-equal competitors and decreasing $B(K)$. The calculation characterizes this breadth–evidence tradeoff rather than optimizing K. In HEIRS, $q = .45$ fixes $K_i = |\mathcal{C}^{(i)}|$ for each history. Screening costs $16K_i$ evaluations; disjoint confirmation adds $T_{\mathrm{conf}}^{(i)}$. History-Cover sets generated at different quantiles need not be nested.

Shared-task variance. For two candidates under the random-probe model,

$$\mathrm{Var}(Y(S) - Y(S')) = \mathrm{Var}(Y(S)) + \mathrm{Var}(Y(S')) - 2\,\mathrm{Cov}(Y(S), Y(S')). \tag{57}$$

Positive covariance lowers the variance of the paired difference relative to the corresponding independent comparison at fixed marginals, whereas negative covariance raises it. This is the classical rationale for common-random-number comparisons Wright & Ramsay (1979). Proposition 6 allows arbitrary within-task dependence because selection is analyzed through paired task-wise differences. HEIRS matches task identities without requiring identical decoding randomness across Skills.

As a descriptive check on the two development panels, 18 of 21 retained-candidate pairs have positive empirical covariance for target 9B (three are negative), while all 21 pairs are positive for target 27B. These are observed panel statistics; the bound above assumes fresh i.i.d. tasks.

Relation to Task-Cover. Task-Cover chooses its 16 tasks deterministically from instruction text, whereas Eqs. 48–54 assume fresh i.i.d. draws. We therefore use the random-task result to explain the role of evidence size, not to certify the HEIRS operating point. On the 4B Spreadsheet development history ($K_i = 7$), the generic bound at $B = 16$ and $\delta = .05$ gives a penalty of about 0.774, which is far too loose to explain the observed gains. The empirical analyses in Section 5.3 and Appendix E provide the relevant evidence for the actual selector.

C Experimental Setup and Baselines

This appendix provides the additional setup needed to interpret the experiments: evaluation conditions, task splits, table conventions, and baseline implementations. Every method chooses its Skill according to its own procedure before evaluation on the common held-out set.

C.1 Evaluation Protocol and Evidence Scope

Development analyses are kept separate from held-out evaluation. Within each comparison, task splits, model settings, and evaluators are fixed before final testing.

GLM evaluation uses the low-reasoning runtime with a 65,536-token output ceiling.

C.2 Task Splits and Main-Table Notes

Source evolution starts from the released Human Skill (SkillOpt `initial.md`); its train/validation splits contain 80/40 tasks on Spreadsheet and 39/18 on ALFWorld. Target Scratch SkillOpt produces `best_skill.md` through its own benchmark-native validation gate, independently of HEIRS selection. When a paired contrast is reported and complete task-level vectors are available, we use an exact paired test together with a paired-bootstrap 95% confidence interval.

Matched-domain ports use the same source split and runtime, and the source-selected Skill is fixed before zero-update target transfer. If two methods produce the same normalized Skill text, we reuse the same target evaluation rather than introduce differences from repeated stochastic execution. In Table 1, dashes denote unavailable or withdrawn results, not zero scores. Unresolved technical outcomes are never imputed.

C.3 Baseline Implementations

Trace2Skill required two technical accommodations that do not change its target-side selection budget. For ALFWorld DeepSeek source synthesis, three incomplete trajectory analyses were completed with a larger output ceiling while the 36 completed analyses were reused. For Spreadsheet Qwen-9B, one source task exceeded the nominal 600-second limit but produced an ordinary hard failure; keeping that failure does not improve the method's score. In the DeepSeek-to-GLM ALFWorld transfer, one episode resumed from its valid prefix after a transport failure while preserving the model, Skill, 50-step episode limit, and 256-token response ceiling. All associated calls are included in resource accounting.

SkillTTA and MASA rows marked * are benchmark/model adaptations whose preserved and omitted components are specified here Wang et al. (2026a); Yu et al. (2026). SkillTTA natively addresses task-

Table 8: **Positioning of cross-executor Skill migration methods.** The distinction is what each method treats as the migration object and what it is allowed to do after the executor changes. This is a scope comparison, not a performance ranking.

Method	Migration question	Object available at migration	Post-change operation	Returned Skill
Direct Endpoint	reuse the source choice?	source-selected Skill	none	same source Skill
WikiSkill Tang et al. (2026)	does an evolved Skill transfer?	evolved persistent Skill	none	same transferred Skill
Target Scratch SkillOpt Yang et al. (2026)	relearn for the target?	initial Skill	target-side optimization	new target-evolved Skill
MASA Yu et al. (2026)	adapt a Skill to the target backbone?	inherited source Skill	target-aware alignment / rewriting	new aligned Skill
SkillTTA Wang et al. (2026a)	synthesize a Skill for the current task?	trajectories + target task	test-time synthesis	temporary task Skill
HEIRS	**which source version should persist?**	**full source Skill history**	**bounded target revaluation**	**unchanged archived Skill**

conditioned test-time synthesis rather than cross-executor inheritance; we include it as an adjacent methodological baseline rather than a same-setting migration method. The port retains trajectory retrieval and task-conditioned temporary Skill synthesis under the same task interface and verifier, with retrieval material restricted to non-confirmation data.

The SkillTTA adapter preserves retrieved-success labels and recorded failure differences. Because it is target-only, one run is shared by the two incoming source directions for the same target executor. It obtains 22/128 (17.2%) on target 4B, 39/128 (30.5%) on target 9B, 54/128 (42.2%) on target 27B, and 23/52 (44.2%) on DeepSeek-to-GLM Spreadsheet.

MASA* rewrites the inherited persistent Skill using target-model feedback. Our adapter uses its general-Skill hill-climbing branch with two one-candidate rounds and accepts a rewrite only when calibration success improves. The task-specific UCB tree search and learned lightweight rewriter are not part of this adapter. The asterisk marks the adaptation from the released Qwen3 implementation to Qwen3.5.

SkillOpt evolves from the common initial Skill. Each target-specific run serves both incoming source directions; optimization and archive-evaluation costs are reported separately.

D Supplementary Transfer Results

This appendix complements Section 5.2 with additional adaptation baselines, paired uncertainty analyses, sample-size robustness, and Qwen9-to-Llama8 selection details. Where an analysis studies the earlier one-stage B16 selector rather than the final two-stage HEIRS selector, we state that distinction directly.

D.1 Additional Target-Side Adaptation Results

Table 9 reports adaptation baselines omitted from the main table. Target-only SkillTTA runs are shared across incoming migrations; source-conditioned MASA-style rewriting retains one result per direction.

D.2 Paired Uncertainty for One-Stage B16 Selection

Table 10 compares the one-stage B16-selected Skill with Direct Endpoint on migrations for which complete paired task outcomes are available. These results quantify the uncertainty of small-probe selection and are separate from the final two-stage HEIRS results. Intervals are percentile paired-task bootstrap intervals, and exact p values use the two-sided binomial form of McNemar's test. Rows with zero discordances and a zero-width interval have identical recorded task outcomes. Comparisons without complete paired vectors are reported only through aggregate success counts. On Spreadsheet 4B→27B, one-stage B16 selection versus SkillOpt has 20/11 discordant tasks over 128 tasks ($+7.03$ pp, 95% interval $[-1.56, +15.63]$ pp, $p = .1496$).

Table 9: **Additional target-side adaptation results.** SkillTTA rows span the two incoming-source columns because they correspond to one physical target-only result per executor. MASA-style Source Rewrite* is source-conditioned and therefore retains directed migration values.

Benchmark	Method	9→4	27→4	4→9	27→9	4→27	9→27	DS→GLM
Spreadsheet	SkillTTA*	17.2		30.5		42.2		44.2
	MASA-style Source Rewrite*	26.6	17.2	32.0	32.8	43.8	64.1	59.6
ALFWorld	SkillTTA*	54.2		43.8		39.6		–
	MASA-style Source Rewrite*	54.2	52.1	54.2	68.8	64.6	79.2	–

Table 10: Paired one-stage B16–Direct contrasts. Differences and intervals are in percentage points; these rows analyze the one-stage selector rather than final two-stage HEIRS.

Domain	Migration	n	B16-only / Direct-only	Difference	95% interval	p
Spreadsheet	9B→4B	128	17/10	$+5.47$	$[-2.34, +13.28]$	0.2478
Spreadsheet	27B→4B	128	6/12	-4.69	$[-10.94, +1.56]$	0.2379
Spreadsheet	4B→27B	128	20/5	$+11.72$	$[+4.69, +19.53]$	0.0041
Spreadsheet	9B→27B	128	0/0	$+0.00$	$[+0.00, +0.00]$	1.0000
ALFWorld	4B→9B	48	10/4	$+12.50$	$[-2.08, +27.08]$	0.1796
ALFWorld	27B→9B	48	6/2	$+8.33$	$[-2.08, +20.83]$	0.2891
ALFWorld	4B→27B	48	0/0	$+0.00$	$[+0.00, +0.00]$	1.0000
ALFWorld	9B→27B	48	0/0	$+0.00$	$[+0.00, +0.00]$	1.0000
Spreadsheet	DS→GLM	52	0/0	$+0.00$	$[+0.00, +0.00]$	1.0000
ALFWorld	DS→GLM	48	0/2	-4.17	$[-10.42, +0.00]$	0.5000

A matched admission comparison on Spreadsheet 27B→9B selects A002 with B16 and A011 with Source Top-K_i. Their held-out scores are 39/128 and 42/128, respectively (-2.34 pp; 17/20 discordances; 95% interval $[-11.72, +7.03]$ pp; $p = .7428$). This comparison isolates the earlier one-stage selection setting.

D.3 Sample-Size Robustness of One-Stage B16 Selection

We also test whether the one-stage B16 results are sensitive to held-out set size. The original 128-task Spreadsheet and 48-task ALFWorld sets are expanded to 200 and 118 tasks with disjoint additional tasks. The selected Skill is held fixed while the evaluation set grows, so this analysis measures result stability rather than reselection.

With the larger evaluation sets, the six-direction mean remains slightly higher for B16 selection on Spreadsheet (34.3% vs. 33.4%) and ALFWorld (58.3% vs. 57.1%). However, all nonzero paired-bootstrap intervals on the added tasks include zero, so the direction-level effects remain uncertain.

D.4 Qwen9-to-Llama8 Spreadsheet Selection Details

The Spreadsheet held-out outcomes are reported in Table 1. HEIRS selects A001 for Qwen9→Llama8 using 176 target evaluations. The initial 16-task confirmation produces a unique leader, so no further confirmation is needed.

Table 11: **Sample-size robustness for one-stage B16 selection.** Task success (%) on the original and expanded held-out sets; the selected Skill is unchanged as the evaluation set grows.

Benchmark	Migration	n_0	Direct$_0$	B16$_0$	n_+	Direct$_+$	B16$_+$
Spreadsheet	9B→4B	128	20.3	25.8	200	20.5	24.5
Spreadsheet	27B→4B	128	20.3	15.6	200	19.0	17.0
Spreadsheet	4B→9B	128	31.3	26.6	200	31.5	25.5
Spreadsheet	27B→9B	128	28.9	30.5	200	30.0	30.5
Spreadsheet	4B→27B	128	43.8	55.5	200	45.5	54.5
Spreadsheet	9B→27B	128	57.0	57.0	200	54.0	54.0
ALFWorld	9B→4B	48	54.2	50.0	118	53.4	49.2
ALFWorld	27B→4B	48	45.8	45.8	118	52.5	52.5
ALFWorld	4B→9B	48	47.9	60.4	118	48.3	56.8
ALFWorld	27B→9B	48	47.9	56.3	118	48.3	51.7
ALFWorld	4B→27B	48	52.1	52.1	118	56.8	56.8
ALFWorld	9B→27B	48	79.2	79.2	118	83.1	83.1

E Admission and Target-Evidence Diagnostics

This appendix separates the two sources of selection error discussed in the main paper: losing a useful Skill during admission and mis-ranking the retained Skills with limited target evidence. We first analyze candidate retention, then screening reliability, and finally alternative ways to spend the same target-evaluation budget. Analyses based on already completed outcome matrices are retrospective and do not change the reported HEIRS results.

E.1 Admission Budget and Candidate Retention

Realized candidate budgets. The History-Cover operating point is fixed at the within-history distance quantile $q = .45$. Table 12 reports the resulting history-specific absolute thresholds and candidate counts K_i; final ordering is determined by target revaluation.

Table 12: Realized candidate budgets under the reported History-Cover quantile $q = .45$. L_i is the byte-distinct source archive size and K_i is induced by the history-specific absolute threshold at that quantile. The reported $16K_i$ quantity is the *Stage-1 screening* cost only; final two-stage HEIRS additionally pays disjoint confirmation cost reported in the main paper.

Benchmark	Source	L_i	K_i	Target calls	Savings vs. full
Spreadsheet	4B	13	7	112	46.2%
Spreadsheet	9B	13	7	112	46.2%
Spreadsheet	27B	13	8	128	38.5%
ALFWorld	4B	5	5	80	0.0%
ALFWorld	9B	9	5	80	44.4%
ALFWorld	27B	4	4	64	0.0%
Spreadsheet (DS→GLM)	DS	13	6	96	53.8%
ALFWorld (DS→GLM)	DS	9	6	96	33.3%

When the archive is already small, the fixed rule admits all states; larger or more redundant histories induce smaller target-evaluation sets. Thus K_i reflects the history's structure at $q = .45$.

Matched admission controls. We compare deterministic and random admission rules at the same seven-state budget.

Table 13: Matched candidate-admission controls on the two Spreadsheet development matrices whose $q = .45$ History-Cover realization yields $K = 7$. Regret is the task-count gap between the best of all 13 states and the best retained state. The unanchored random control averages over 1,716 seven-state subsets; its endpoint-anchored counterpart averages over 924 subsets containing A001.

Candidate rule	K	Regret 4B→9B	Regret 4B→27B
Full history	13	0.000	0.000
History-Cover ($q = .45$)	7	0.000	0.000
Source Top-K_i	7	0.000	0.000
Archive-order uniform (unanchored)	7	0.000	0.000
Random same-K (unanchored)	7	0.283	0.944
Random same-K + endpoint	7	0.318	1.030

These local matched-K_i controls show finite-panel retention at lower candidate count. The deterministic admission rules produce the same outcome on the two development matrices, while endpoint-anchored random subsets incur nonzero mean regret. The unanchored rows serve as descriptive subset controls.

Figure 6(a–b) in the main paper gives the broader complete-panel comparison. That aggregate measures multi-domain retention and candidate regret, while the development controls above isolate admission behavior on the development histories.

E.2 Screening Fidelity and Probe Budget

Probe-to-held-out rank fidelity. To test whether a 16-task screen preserves the ranking beyond its observed tasks, we evaluate all 13 archived Skills on a separate 52-task DeepSeek-to-GLM Spreadsheet panel. The 16-task set is the Stage-1 screen used by HEIRS; the larger panel is used only to measure ranking fidelity and does not affect screening or confirmation.

Table 14: **Probe-to-held-out historical ranking fidelity.** We compare the 16-task screening order with a separate 52-task ranking over the same historical Skills. Inv. is the fraction of comparable Skill pairs whose order reverses.

State set	Spearman ρ	Kendall τ_b	Inv.	Selected regret
Full history (13)	0.534	0.415	27.0%	11.54 pp
History-Cover admitted (6)	0.638	0.552	21.4%	11.54 pp

On the full archive, B16 gives the highest score to A007 and A010 (13/16), while A009 has the highest point estimate on the 52-task panel (37/52). The B16 tie-break selects A007, which reaches 31/52. A009 wins eight discordant tasks and A007 wins two; exact McNemar $p = .109$ and the paired-bootstrap 95% interval is $[0.0, 23.1]$ pp. The larger panel therefore shows substantial ranking uncertainty but does not establish that A009 is population-optimal. Among the six admitted Skills, source validation and B16 have similar Spearman correlation with the 52-task ranking (0.657 and 0.638), reinforcing that limited target evidence can still mis-rank retained candidates.

Probe-size sensitivity. Table 15 varies the number of common tasks used to rank the same seven candidates.

Table 15: Selection quality as the common-task budget increases for the seven-candidate development sets. Regret counts lost successes; recovery is relative to the endpoint-to-best gain on all 16 tasks.

	4B→9B		4B→27B	
B	Regret	Recovery	Regret	Recovery
4	0.926	0.537	1.901	0.620
8	0.451	0.774	0.366	0.927
12	0.108	0.946	0.020	0.996
16	0.000	1.000	0.000	1.000

At $B = 16$, the probe is the complete diagnostic panel and selector regret on that panel is mechanically zero; separate held-out evaluation measures generalization beyond the probe.

Table 12 varies candidate coverage; Table 15 instead varies the evidence used to rank a fixed candidate set.

A single realized Task-Cover prefix need not improve monotonically with additional tasks: on this lane the full-history Spearman correlation is 0.664 at $B = 12$ and 0.534 at $B = 16$. Thus the $B = 16$ operating point should not be interpreted as a guarantee of deployment ranking; Figure 7(b) reports the aggregate disjoint-split budget trend.

E.3 Alternative Probes and Independent Confirmation

Alternative fixed-budget probe construction. We next ask whether a different choice of 16 target tasks would preserve candidate rankings better. Semantic coverage groups tasks by instruction similarity; an alternative is to group tasks by whether historical Skills succeed and fail on them in the same pattern. Following the response-structure perspective of Anchor Points and tinyBenchmarks Vivek et al. (2024); Maia Polo et al. (2024), we define the behavior signature of task x as

$$z(x) = \big(R(S_1, x), \ldots, R(S_M, x)\big)$$

from a panel of previously evaluated development Skills. Tasks with identical signatures form a behavior stratum. BAP allocates at least one representative to every stratum, assigns remaining budget to strata with the largest uncovered mass, selects extra representatives inside a stratum by semantic medoid plus farthest-first coverage, and weights each selected anchor by the number of calibration tasks it represents. The target-side score is the weighted mean over the selected anchors. Identical within-stratum signatures ensure exact full-pool scores for all development states.

For the 24-task Spreadsheet development pool, the eight previously evaluated Qwen Skill states induce only 11 distinct behavior signatures: one 14-task all-zero stratum and ten singleton strata. BAP-16 therefore keeps all ten discriminative singleton patterns and allocates six representatives inside the large redundant stratum. The resulting weighted 16-task probe exactly reproduces the full-24 development hard counts $[4, 5, 5, 6, 3, 2, 5, 4]$ for all eight development states.

Table 16: **Alternative 16-task probe constructions.** Each method selects 16 tasks, and we compare the resulting Skill ranking with the separate 52-task ranking over the six admitted Skills. "Regret" is the 52-task point-estimate gap to the best Skill in that set.

Probe construction	Spearman ρ	Selected	Regret (pp)
Current Task-Cover	0.638	A007	11.54
Semantic Facility Location	0.657	A007	11.54
Semantic DPP/log-det	0.657	A007	11.54
Historical item discrimination	0.696	A007	11.54
Behavior-Anchored Probe	**0.771**	**A009**	**0.00**

The result is stable across several weighting choices: all tested weights select A009 with $\rho = .771$ for the same 16 anchors. Exhaustively enumerating all $\binom{24}{16} = 735{,}471$ subsets shows that 16 tasks can support substantially higher rank fidelity on this matrix: the median subset has $\rho = .736$, the best reaches .971, and 23.2% select A009. The limitation is therefore not the budget size alone; task composition also matters. These results do not identify a generally optimal 16-task selector.

As a secondary fixed-budget analysis, Behavior-Anchored Probe (BAP) replaces semantic task coverage with source-side behavior strata. We apply the same 16-anchor coreset to two additional Spreadsheet migrations; it exactly reproduces the 40-task source-validation scores of all 12 measured source states and is fixed before either target-side comparison.

Table 17: **Behavior-Anchored Probe on two additional migrations.** A source-exact 16-anchor coreset selects a Skill for each migration; held-out results are compared with the corresponding B16-selected reference Skill.

Migration	BAP Skill	BAP evidence	Reference Skill	Reference evidence
4B→27B	A011	$\leq 70/128$	A012	71/128
4B→9B	A010	$\geq 37/128$	A012	34/128

The two migrations give different target-side orderings. For 4B→27B, BAP selects A011, whose held-out upper bound is below the A012 reference; for 4B→9B, BAP selects A010, whose held-out lower bound exceeds the A012 reference. Exact reconstruction of source-side validation behavior therefore does not preserve target ranking consistently across executors. BAP is reported as a secondary probe-construction analysis rather than as the primary selector.

Leave-one-executor-out behavior coresets. A stronger source-only variant builds the 16-task coreset from behavior under more than one previous executor while excluding the target executor. This broader source view still does not reliably preserve target ranking. On 9B→4B, it ranks A004 above A012 (14 vs. 12), yet A004 can reach at most 25/128 held-out successes compared with 33/128 for A012. On 4B→9B, it ranks A011 above A006 (10.67 vs. 9.83), while A011 reaches at least 39/128 on the held-out set. These results show that richer source-side behavior is informative but still insufficient for final target-side selection.

Independent confirmation after screening. A separate two-state experiment tests the value of using fresh evidence for final commitment. Two Skills are shortlisted by one target set and then compared on a disjoint confirmation set; the screening scores are not reused. This experiment isolates the benefit of sample splitting, while Table 1 reports the full HEIRS procedure.

Table 18: **Sample-split confirmation analysis.** Screening and confirmation use disjoint target-task sets, and confirmation compares only the two shortlisted Skills. The 128-task held-out set provides an external check.

Migration	Screen	Confirm	Commit	Held-out 128
9B→4B	A004 14 / A012 12	A004 1 / A012 2	A012	A012 33; A004 $\leq$25
4B→9B	A011 10.67 / A006 9.83	A011 2 / A006 4	A006	A006 $\geq$40; B16 34

The disjoint confirmation set reverses the screening choice in both migrations. On 9B→4B it selects A012, which scores 33/128 on the held-out set, while A004 has an optimistic upper bound of 25/128. On 4B→9B it selects A006, which reaches at least 40/128. These cases show that fresh confirmation evidence can correct a screening choice that does not generalize.

Task-support coverage at fixed B. At fixed target budget, selection error also depends on *which* target tasks are observed. We test a simpler outcome-independent representative probe: concatenate normalized word TF–IDF $(1, 2)$ and character `char_wb` TF–IDF $(3, 5)$ features of target-task instructions, then select 16 common tasks by deterministic medoid plus farthest-first coverage. We compare this static representative B16 rule with a posterior-weighted 8+8 adaptive discrimination rule on four already-complete Spreadsheet candidate–task matrices.

Table 19: **Task selection at a fixed 16-evaluation budget.** The static representative probe chooses all tasks from instruction text, whereas the 8+8 rule chooses its second block adaptively. "Oracle" is the best Skill on the complete matrix.

Matrix	K	Oracle	Rep. B16	Regret	Adaptive 8+8	Regret
DS→GLM held-out52	6	A009	A009	0.00	A009	0.00
H4→9B held-out48	3	A006	A003	2.08	A003	2.08
H9→4B held-out48	4	A005	A005	0.00	A005	0.00
H9→9B held-out48	4	A005	A005	0.00	A005	0.00

Across these four matrices, the static and adaptive rules select the same Skill in every case. The static probe matches the complete-matrix best Skill in three cases and misses by one task (2.08 pp) in the fourth. Together with the exhaustive subset analysis above, this shows that which tasks are observed can be as important as the number of tasks: an unrepresentative small set can distort the margins between candidates.

E.4 Adaptive Evidence and Full-History Selection

Adaptive evidence allocation. Adaptive Evidence Extension (AEE) addresses probe uncertainty by allocating more tasks to near-equal candidates. Starting from B16, it retains candidates within one hard-success unit of the leader and gives them additional common-task blocks. Task blocks follow $16 + 8 + 4 + 4$, pooling scores over each surviving candidate's observed tasks. B16 score resolution fixes the one-task ambiguity band before held-out evaluation.

Table 20: **Adaptive evidence versus fixed probe budgets.** Means use 50,000 random task orderings of the complete six-candidate cross-family matrix. All policies reserve up to 32 evidence positions and use the same 20 held-out tasks. AEE gives extra evidence to near-equal candidates.

Policy	Mean calls	Mean regret (pp)	Top-set hit (%)	Harm vs. endpoint (%)
Fixed B16	96.0	6.33	54.3	12.4
AEE $16 + 8 + 4 + 4$	**118.8**	**4.27**	**68.4**	**10.4**
Fixed B24	144.0	4.96	64.4	10.7
Fixed B32	192.0	3.20	74.6	7.2

In resampling, AEE uses fewer target rollouts than fixed B24 while attaining lower mean held-out regret and a higher top-set hit rate. On the realized cross-family matrix, B16 leaves A007 (13/16) and A009 (12/16) within the ambiguity band; twelve additional common tasks move the cumulative scores to 20/28 and 21/28, respectively, so AEE would select A009 after 120 logical rollouts rather than evaluating all six candidates to B24 (144 calls). AEE is reported as a secondary mitigation analysis and does not modify the reported HEIRS method.

Progressive full-history ablation. A q-free variant gives every archived state initial target evidence and progressively halves candidates under a fixed logical-call ceiling. On a disjoint 72-task Spreadsheet extension for 4B→9B, Direct A001 reaches $23/72 = 31.94\%$ while progressive revaluation selects A004 at $19/72 = 26.39\%$, a -5.56 pp difference with paired-bootstrap 95% interval $[-15.28, +4.17]$ pp. This illustrates the breadth–evidence tradeoff: full admission avoids omitting candidates, but a fixed budget spread across more competitors can worsen selection.

E.5 Checks of Selection Bounds

Theory-facing checks. We test the distortion bound and the variance reduction from task pairing.

Table 21: Selector stability on the same development matrices. "Cert." reports the fraction of 10,000 random common-task prefixes satisfying $\Omega_{\mathcal{C}} < \gamma_{\mathcal{C}}$; "Best hit" reports the fraction selecting a retained state with maximal utility on the complete B16 panel.

	4B→9B		4B→27B	
B	Cert. (%)	Best hit (%)	Cert. (%)	Best hit (%)
4	0.0	58.3	3.1	49.0
8	1.5	77.7	16.2	88.7
12	46.1	94.6	88.7	99.3
16	100.0	100.0	100.0	100.0

For all prefixes satisfying the sufficient stability condition, the selected state is complete-panel optimal, as required by Corollary 3. Equation 38 holds for all 60,000 nontrivial prefixes ($B < 16$).

Table 22: Effect of evaluating retained candidates on shared task identities. The variance ratio is paired-difference variance divided by the counterfactual sum of marginal variances under independent task assignment. Values below one indicate lower variance under pairing.

Target	Positive covariance	Lower paired variance	Median variance ratio
9B	18/21	18/21	0.56
27B	21/21	21/21	0.41

F Migration Costs and Reproducibility

This appendix gives the detailed cost breakdown behind Section 5.4 and summarizes the materials needed to reproduce the experiments.

F.1 Migration Cost Details

Source evolution happens before migration and is reused by HEIRS rather than repeated. History-Cover only compares archived Skill texts and requires no target-model inference; a complete pairwise scan needs $\binom{L_i}{2}$ text comparisons (78 for 13 Skills and 36 for 9). We therefore separate the cost of selecting from an existing archive from the cost of learning a new target Skill.

Archive-only screening. For the cross-family costs not tabulated in the main paper, HEIRS Stage 1 costs $16K_i$ target calls after History-Cover admission, whereas exhaustive Full History screening costs $16L_i$.

Table 23: **Cross-family archive-screening cost.** B16 target screening evaluations for the DeepSeek→GLM migrations. HEIRS uses $16K_i$ Stage-1 calls and Full History uses $16L_i$ calls; HEIRS confirmation is excluded from both columns.

Benchmark	HEIRS Stage 1	Full History B16	Fewer calls
Spreadsheet (DS→GLM)	96	208	53.8%
ALFWorld (DS→GLM)	96	144	33.3%

Fresh target relearning. Target Scratch SkillOpt starts from the common initial Skill and learns a new Skill for each target executor. One target-only Scratch run serves the two incoming source directions for that executor. These runs use 920 logical rollouts per Spreadsheet target and 348/420/330 rollouts for ALFWorld targets 4B/9B/27B. Their native-token totals are Spreadsheet: > 38.527M, 23.717M, and 11.398M for 4B/9B/27B; ALFWorld: 18.741M, 23.202M, and 19.142M. SkillTTA and MASA use different synthesis or rewriting operations, so we do not combine their costs with rollout counts. HEIRS token coverage is incomplete at the candidate level; we therefore compare methods using logical target evaluations and report other resource measures separately.

F.2 Reproducibility Artifacts

The supplementary artifacts include the exact Skill files, source histories, task manifests, per-task score matrices, q sweeps, random-control distributions, statistical procedures, and detailed resource records needed to reproduce the reported results.