FROM MULTI-AGENT TO SINGLE-AGENT: WHEN IS SKILL DISTILLATION BENEFICIAL?

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent systems (MAS) for structured data-science tasks externalize analytical control through workflows, but repeatedly executing orchestration incurs deployment overhead. Distilling workflows into single-agent skills raises a selective-transfer question: which components should survive? We distinguish capability resources, which expand available actions and knowledge, from pipeline guidance, which constrains explored solutions. On the same causal-estimation instances, capability-matched pipeline retention changes normalized utility by $+20.0$ points under method-selection accuracy but -10.3 under numerical error. We introduce Behavior-Outcome Freedom (F), a pre-synthesis diagnostic of signed behavior–outcome rank mismatch, and give a candidate-conditional account through Signed Anchor-Rank Transfer with an explicit remainder for coarse representations. AdaSkill preserves validated capabilities, removes runtime orchestration, and calibrates an F-threshold on module-level development interventions, freezing it before target construction. Across 16 matched interventions, structural effect decreases with F ($r = -0.80$, $p < 0.001$); 15 atomic treatments localize the reversal to pipeline guidance. Across 11 datasets and four task families, directly executed skills combine strong performance with lower latency than source workflows. Workflow-excluded construction tests the calibrated policy separately from the mechanism analysis.

1 INTRODUCTION

Multi-agent systems (MAS) for structured data-science tasks use workflows to organize analytical control across roles, stages, shared state, tool calls, verification, and repair (Cao et al., 2026; Verma et al., 2025; Shen et al., 2025; Ouyang et al., 2025). These workflows achieve strong performance but externalize control into runtime orchestration that must be repeatedly executed at deployment. Skill distillation offers an alternative: transfer reusable workflow competence into a single tool-using agent, reducing coordination overhead while preserving source-system analytical competence.

Existing approaches pursue two complementary forms of agent distillation. MAS-distillation methods compress interaction graphs, debate dynamics, or execution trajectories into a single student model (Chen et al., 2024; Aluru et al., 2025; Luo et al., 2026), while inference-time methods synthesize, verify, attribute, or refine reusable skills from source workflows and execution experience (Qiu et al., 2025; Alzubi et al., 2026; Ma et al., 2026; Yuan et al., 2026). These methods internalize, construct, evaluate, or optimize transferred agent logic. It remains unclear whether functionally different source components should cross the control boundary under the same inheritance rule.

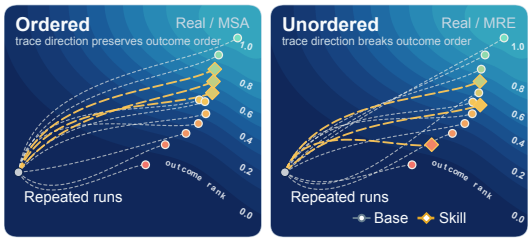


Figure 1: **Behavior–outcome geometry governs structural lift.** Repeated runs share an input; trajectories schematically encode behavior and endpoints encode outcome rank. **Left:** aligned distance ranks support structural guidance. **Right:** broken ordering favors flexibility.

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.

However, a MAS workflow does not transfer a functionally homogeneous object. Callable tools and declarative knowledge are *capability resources* that expand what a target agent can access, whereas ordering rules, decision gates, and other *pipeline guidance* impose source-derived control over which reachable solutions the agent explores. As illustrated in Figure 1, when a task-qualified pipeline contracts behavior toward successful anchors, positively ordered behavior–outcome geometry allows that contraction to carry outcome-rank information; as the ordering weakens or reverses, the same source control loses reliable support and can suppress valuable alternatives. Thus, even a published, high-performing workflow need not provide beneficial structural control after distillation. Our central question is: *when does transferring workflow structure from a multi-agent system benefit a single-agent skill?*

Structural transfer value depends jointly on behavior–outcome geometry and on the strength and direction of the transferred intervention. We therefore distinguish *capability transfer*, which preserves tools and declarative knowledge, from *structural transfer*, which preserves source-derived pipeline guidance. To characterize the upstream geometry before skill synthesis, we introduce **Behavior-Outcome Freedom** (F), the signed rank mismatch between pairwise behavioral distance and score distance. Low F indicates positively ordered geometry in which behavioral contraction can faithfully carry outcome-rank information; as F approaches one, monotone organization weakens, and above one the geometry becomes anti-concordant. Importantly, F is an effect modifier rather than a stand-alone estimator of candidate lift: a task-qualified source workflow supplies a grounded structural proposal, while the realized strength and direction of that proposal remain candidate-specific. Our Signed Anchor-Rank Transfer analysis formalizes this candidate-conditional relationship. For coarse behavioral representations, an explicit remainder retains execution-quality changes hidden within a behavior label, instead of assuming an unchanged outcome kernel.

We propose **AdaSkill**, an adaptive workflow-to-skill distillation policy that preserves validated capability resources, removes runtime orchestration, and makes pipeline guidance the sole adaptive object. AdaSkill decomposes a task-qualified source workflow into callable tools, declarative knowledge, pipeline guidance, and runtime orchestration; retains validated tools and knowledge; removes orchestration unnecessary for single-agent execution; and applies a development-calibrated rule, frozen before target evaluation, that uses F to determine whether source-pipeline guidance should be inherited. For deployment, module-level sign calibration learns the threshold and attains empirical value optimality on the observed interventions, separately from the candidate-conditional mechanism theorem. We study this mechanism through scorer-controlled interventions, atomic Pipeline-only treatments, capability-matched Full-versus-Discard comparisons, and workflow-excluded construction across 11 datasets spanning text-to-SQL, causal estimation, causal discovery, and feature engineering. The skills combine strong performance with lower latency than their source workflows. Our contributions are:

- **Functional Decomposition of Workflow Transfer:** We distinguish capability transfer from structural transfer and show why tools and declarative knowledge should not be governed by the same inheritance rule as source-pipeline control. This decomposition turns workflow distillation from monolithic compression into a selective transfer problem.
- **Behavior–Outcome Mechanism and Diagnostic:** We introduce Behavior-Outcome Freedom (F) as a pre-synthesis diagnostic of signed behavior–outcome geometry and provide a candidate-conditional theoretical account through Signed Anchor-Rank Transfer and the Structural-Lift Decomposition. Scorer reversal, 15 atomic Pipeline-only treatments, and 16 capability-matched Full-versus-Discard interventions show structural value declining with F.
- **Adaptive Single-Agent Skill Distillation:** We develop AdaSkill as a constructive policy that preserves validated capability resources, removes runtime orchestration, and conditionally inherits source-pipeline guidance using a calibrated rule frozen before target evaluation. Across 11 datasets spanning program construction, numerical estimation, graph recovery, and predictive modeling, AdaSkill combines strong performance with lower source-workflow latency.

2 Background and Related Work

Workflow control and distillation. AutoGen and MetaGPT externalize control through explicit interactions (Wu et al., 2024; Hong et al., 2024); data-science workflows specialize it into SQL, estimation, graph search, and feature engineering (Cao et al., 2026; Verma et al., 2025; Shen et al.,

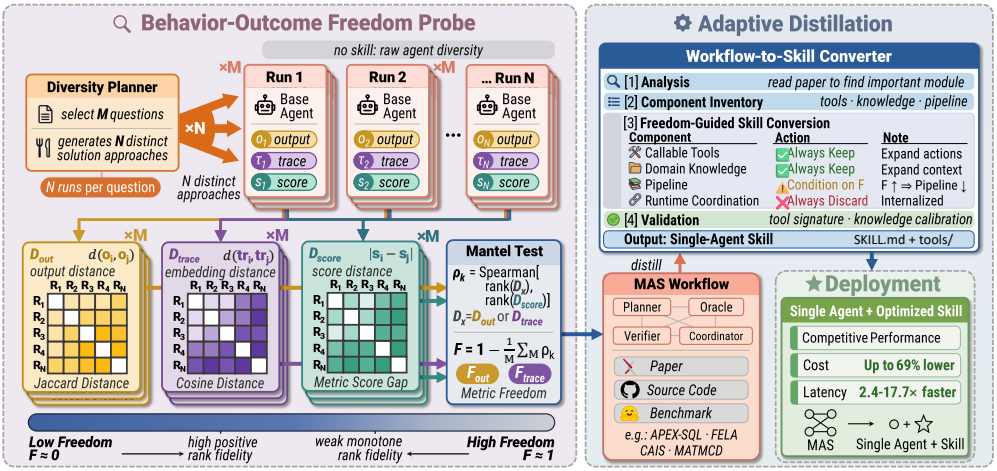


Figure 2: **AdaSkill operationalizes signed structural transfer.** The converter uses pre-synthesis F to condition pipeline inheritance, preserves validated capabilities, removes orchestration, and emits a tool-using skill. Deployment figures compare task-average execution costs and latency with source workflows (Table 1); CE is the cost exception.

2025; Ouyang et al., 2025). Single-agent execution and context management internalize control (Li, 2026; Xu et al., 2026c;a;b); AgentArk, SMAGDI, Latent Agents, and MAPD compress source protocols (Luo et al., 2026; Aluru et al., 2025; Yi et al., 2026; Liu et al., 2026). AdaSkill studies transfer selectivity rather than compression fidelity.

Skill optimization. Voyager, EvoSkill, Trace2Skill, and ASDA develop or adapt skills from experience (Wang et al., 2023; Alzubi et al., 2026; Ni et al., 2026; Yim et al., 2026); AutoSkill, SkillRL, and Memento-Skills evolve libraries (Yang et al., 2026b; Xia et al., 2026; Zhou et al., 2026). SkillGen verifies effects, ClawTrace attributes cost, SkillOpt optimizes text, and SKILL-DISCO compiles shared traces (Ma et al., 2026; Yuan et al., 2026; Yang et al., 2026a; Guo et al., 2026). These methods optimize candidates or trajectories. AdaSkill diagnoses geometry before synthesis, separates resources from directives, and calibrates retention through controlled interventions.

3 Methodology

3.1 Selective Workflow Transfer

Capability resources enable actions; pipeline directives select among them. Capability text need not leave behavior unchanged; both candidates remove orchestration. A *task-qualified* source is published and high-performing for the same objective, maps its clauses to scored decisions, and passes target-interface checks. Qualification supplies a grounded proposal, not a positive-transfer guarantee (Figure 2).

3.2 Behavior-Outcome Freedom

Let H be the complete evaluator-facing execution state and $B = \phi(H)$ its diagnostic projection. The evaluator scores H; B need not be outcome-sufficient. For example, estimator identity omits covariates and calls that affect numerical error. Within-B changes enter an explicit remainder.

Estimand and estimator. Fix an evaluation unit q and two independent baseline runs $Z, Z' \sim \mathbb{P}_{0,q}$ under agent A and harness $\mathcal{H}$. For symmetric behavioral dissimilarity d and score $s \in [0,1]$, set $U = d(\phi(Z), \phi(Z'))$ and $V = |s(Z) - s(Z')|$. Define

$$F_q = 1 - \rho_S(U, V), \qquad \rho_S(U,V) = \mathrm{Corr}(R_U(U), R_V(V)), \tag{1}$$

where $R_U(u) = \Pr(U < u) + \frac{1}{2}\Pr(U = u)$ handles ties, and similarly for V. Both rank variances must be nonzero. For n observed runs, rank the upper-triangular behavioral and score-distance entries using average ranks:

$$r_M(\mathcal{X}) = \mathrm{Corr}_{i<j}(R^{\mathcal{X}}_{ij}, R^{\mathrm{score}}_{ij}), \qquad F_{\mathcal{X}} = 1 - r_M(\mathcal{X}). \tag{2}$$

This Mantel-style statistic (Mantel, 1967) is scale-invariant. $F \in [0, 2]$: values near zero indicate positive rank coupling, near one weak signed ordering, and above one anti-concordance, which reverses the rank implication of contraction. None alone identifies native-scale candidate benefit.

Representation, aggregation, and routing. Output-space F_{out} uses task-authored structured behavior and dissimilarities (Appendix Table 3). Trace-space F_{trace} uses intermediate messages, tool calls, and results with the final response removed, embedded by `text-embedding-3-large` (OpenAI, 2024); distance is $(1 - \cos_{\text{sim}})/2$. A CE audit additionally removes selected-method fields and enriches output behavior with response-free execution signatures, without outcome values.

For question-level tasks, average F_q over identifiable questions; dataset-level tasks use the pooled unit. The 16 dataset–metric intervention coordinates differ from the seven converter inputs, which aggregate pre-specified diagnostic datasets or use a designated conversion benchmark. Per-question routing requires at least three identifiable questions. Output F is primary; trace F is a fallback only when output is unidentifiable. If neither qualifies after additional sampling, use Discard. Constant distance vectors are undefined, not zero; complete descriptive and strict-identifiability analyses are reported separately.

Sampling and uncertainty. A diversity planner seeds N Base Agent runs on each of M diagnostic questions with distinct plausible strategies. This is a planner-design diagnostic, not an iid estimate of Eq. (1); the distinction is formalized in Appendix C. Independent-run controls test dependence on that law (Section 4.6). Shared-run pairs are dependent: uncertainty resamples runs or questions, never individual pairs. Finite-sample screening for observed score variation is part of the protocol (Appendix C). For reporting, normalized lift is $(u_{\text{skill}} - u_{\text{base}})/(1 - u_{\text{base}})$ for higher-is-better utility, or $(e_{\text{base}} - e_{\text{skill}})/e_{\text{base}}$ for MRE.

3.3 Formal Account of Structural Transfer

At representation B, baseline and candidate laws factor as $P_0 = \nu\mu_0 K_0$ and $P_\pi = \nu\mu_\pi K_\pi$, with behavior law μ and conditional score law K. Let $m_j = \mathbb{E}_{K_j}[s \mid q, B]$. Under behavioral overlap,

$$\text{Lift}(\pi) = \underbrace{a_\pi S_\pi G}_{T_B(\pi)} + \underbrace{\mathbb{E}_{\nu\mu_\pi}[m_\pi - m_0]}_{R_{\pi,B}}. \tag{3}$$

G is baseline outcome relief, S_π induced shift, and a_π directional alignment. The exact remainder $R_{\pi,B}$ retains changes hidden by coarse B; it vanishes for evaluator-sufficient B, not by assumption for estimator identity. Overlap is assumed, not established by interface checks. Appendix A defines the factors and proves the decomposition and supported-mixture bound.

Enrichment audits diagnostic stability, not $R_{\pi,B} = 0$; matched outcomes measure lift.

If Γ_{rank} is score-distance rank variation explained by behavior distance,

$$\rho_S(U, V)^2 \le \Gamma_{\text{rank}}, \qquad 1 - \Gamma_{\text{rank}} \le 2F - F^2. \tag{4}$$

The anchor argument applies to the virtual behavior-only law $\widetilde{P}_\pi = \nu\mu_\pi K_0$, with finite $\chi^2_{\pi,B} = \mathbb{E}_0(w_\pi - 1)^2$, not directly to a kernel-shifting candidate. Let Δ_π denote standardized behavior-rank contraction toward a successful Base event of mass $h > 0$, and Λ_π the corresponding score-rank contraction under this virtual law. Signed Anchor-Rank Transfer gives, for $\kappa \in \{-1, +1\}$,

$$|\Lambda_\pi - \kappa\Delta_\pi| \le \sqrt{2\chi^2_{\pi,B}\{1 - \kappa(1 - F)\}/h}. \tag{5}$$

Positive-orientation error scales as $\sqrt{2\chi^2_{\pi,B}F/h}$; near anti-concordance the direction reverses. This is a *candidate-conditional mechanism*, not a threshold derivation: candidates sharing F can have opposite alignment. $\Delta_\pi, \chi^2_{\pi,B}, h$ are not converter inputs. Appendix A gives complete assumptions, question-mean bounds, and impossibility boundaries.

Contractions use centered, variance-one population mid-ranks. Native-score lift additionally requires a rank-to-score link and the representation remainder; rank contraction alone is insufficient.

Operational policy bridge. The implemented module-level sign calibration also attains empirical value optimality on these data. For pre-synthesis F_i, capability-matched Retain/Discard utilities

Table 1: **Dataset-level performance and domain-level efficiency across four task families.** Performance is reported in percent; T2SQL cells list EX, NSF, and SRR in that order. Latency and cost are averaged across datasets within each task family. Boldface denotes the best displayed result, including ties. Taobao component AUCs are recovered from Figure 5.

Dataset	Metric(s)	Source Workflow	Base Agent	EvoSkill	Workflow Compiler	*Tools*	*Knowledge*	*Pipeline*	AdaSkill
Causal Estimation									
Textbook	MSA ↑ / MRE ↓	83.3/54.0	61.5/31.2	71.8/25.4	82.1/26.3	69.2/28.3	79.5/25.6	84.6/26.5	**89.7/23.1**
Synthetic	MSA ↑ / MRE ↓	75.9/16.2	**100.0**/10.6	**100.0**/7.1	**100.0**/12.9	95.6/5.5	91.1/6.0	**100.0**/5.5	**100.0/4.3**
Real	MSA ↑ / MRE ↓	78.3/32.0	87.1/18.3	93.5/21.6	90.3/17.3	87.1/17.0	90.3/18.0	**96.8**/17.8	**96.8/15.9**
Latency (min) / Cost ($) ↓		2.1/**.172**	**0.6**/.296	1.1/.270	1.9/.608	0.8/.208	0.9/.305	0.9/.295	0.9/.207
Causal Discovery									
AutoMPG	F1 ↑	**73.7**	67.0	70.0	70.0	65.0	67.0	67.0	65.2
DWDClimate	F1 ↑	66.3	55.8	**74.0**	64.0	54.0	**74.0**	61.0	70.1
Sachs	F1 ↑	55.7	86.9	87.0	67.0	89.0	89.0	**95.0**	92.4
Child	F1 ↑	65.3	80.3	90.0	86.0	85.0	79.0	81.0	**95.2**
Latency (min) / Cost ($) ↓		32.5/3.01	3.0/2.22	4.2/2.68	2.8/1.58	1.8/**.43**	1.1/.48	**0.8**/.52	3.1/2.50
Feature Engineering									
Taobao	AUC ↑	65.3	66.0	66.4	65.8	65.9	66.7	65.4	**67.0**
Dia	AUC ↑	**81.2**	80.4	80.6	79.8	81.1	81.0	80.3	**81.2**
Latency (min) / Cost ($) ↓		630.0/13.63	44.0/5.11	72.3/6.97	42.9/**2.78**	73.8/7.33	70.4/7.04	89.4/9.64	**35.6**/4.26
Text-to-SQL									
BIRD-147	EX / NSF / SRR ↑	**69.4**/60.8/**93.9**	59.2/90.5/71.8	66.0/93.0/76.2	67.4/93.7/79.6	66.0/92.9/75.5	66.7/93.1/74.2	64.6/91.7/68.7	67.4/**94.1**/80.8
Spider-120	EX / NSF / SRR ↑	**67.5**/47.0/**88.3**	60.0/11.4/0.8	62.5/20.8/2.5	57.5/13.4/1.7	63.3/28.5/3.3	64.2/45.3/5.0	60.8/16.2/1.7	66.7/**60.5**/7.5
Latency (min) / Cost ($) ↓		6.0/.72	**1.3**/.39	1.6/.48	2.4/.59	1.7/**.37**	1.6/.51	2.1/.88	1.7/.45

u_i^R, u_i^D, effects $\delta_i = u_i^R - u_i^D$, and pre-specified cluster weights ω_i, empirical policy learning (Athey & Wager, 2021) selects

$$\widehat{t} \in \arg\max_{t \in \mathbb{R}} \widehat{V}(t), \qquad \widehat{V}(t) = \sum_i \omega_i \{u_i^D + \mathbf{1}(F_i < t)\delta_i\}. \tag{6}$$

Proposition 3.1 (Threshold-policy bridge)**.** *For population value $V(t)$ and $t^\star \in \arg\max_t V(t)$, $V(t^\star) - V(\widehat{t}) \le 2\sup_t |\widehat{V}(t) - V(t)|$. If $\tau(f) = \mathbb{E}[u^R - u^D \mid F = f]$ is non-increasing and F has no atom at its sign-change boundary, the class contains a Bayes-optimal F-only policy.*

The guarantee is best-in-class average value, not individual safety; monotonicity is assumed, not derived from the anchor theorem. Four held-out regimes do not establish a small population uniform error.

3.4 Adaptive Workflow-to-Skill Converter

Inventory and validation. The converter retains target-valid tools and declarative knowledge, classifies all state-to-action and mixed clauses as Pipeline, and removes orchestration. A clause audit rejects source directives in Discard and checks imports, interfaces, descriptive tone, and grounding (Appendix C.2).

Calibrate, freeze, and execute. Each fold selects $\widehat{t}$ on other regimes, retaining pipeline iff $F < \widehat{t}$. Thresholds span 0.426–0.692; 0.644 is the rounded pooled reference. Calibration minimizes module-level sign errors; here perfect separation also makes its solution value-optimal (Appendix A.10). Single-metric families execute the selected candidate; multi-metric families execute a jointly composed skill. Two global endpoints cannot identify module interactions or guarantee hybrid optimality.

Both candidates share deduplicated, validated resources. The emitted skill is re-audited without manual relabeling; knowledge headings do not exempt directives.

4 Experiments

4.1 Setup

Tasks and systems. We evaluate 11 datasets: BIRD-147 and Spider-120 for Text-to-SQL (Li et al., 2023; Lei et al., 2025); three CauSciBench datasets for Causal Estimation (CE) (Acharya et al.,

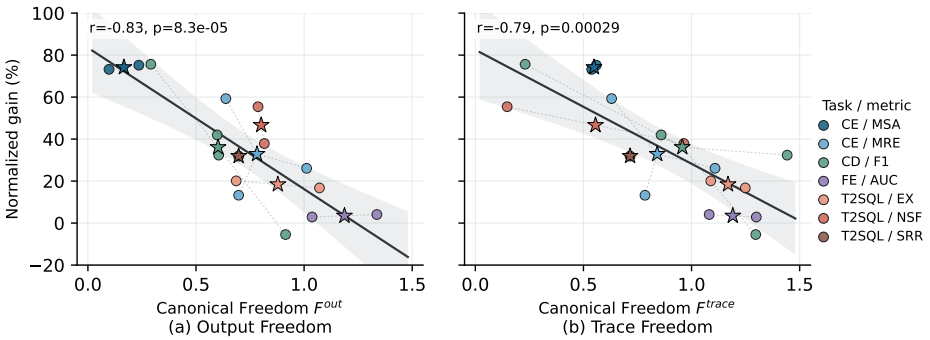


Figure 3: **Overall skill utility follows Behavior-Outcome Freedom.** Output **(a)** and response-free trace **(b)** geometry recover the same ordering across 16 dataset–metric tuples. Circles: datasets; stars: metric means. Lift includes capability transfer; Figure 6 isolates structure.

2026); AutoMPG, DWDClimate, Sachs, and Child for Causal Discovery (CD) (Quinlan, 1993; Mooij et al., 2016; Sachs et al., 2005; Spiegelhalter, 1992); and Taobao and Dia for Feature Engineering (FE) (Alibaba Group, 2018; Teboul, 2022). Source workflows are in Appendix Table 2. Base Agent, Workflow Compiler (Li, 2026), EvoSkill (Alzubi et al., 2026), and AdaSkill share Sonnet 4.6, tools, splits, and evaluator; native Source Workflows serve as deployment references.

Protocol. CE uses MSA/MRE; T2SQL uses execution accuracy (EX) and schema metrics; CD uses edge F1; FE uses AUC. Diagnosis, conversion, and final evaluation have disjoint roles. Gold scores enter diagnostic distance ranks; the converter sees metric-level F, not instances or answers. Development thresholds are frozen before target execution. RQ1 changes the scorer, RQ2 isolates pipeline guidance through capability-matched interventions, and RQ3 tests excluded-workflow construction (Appendix B).

4.2 End-to-End Results

Performance and efficiency. AdaSkill combines competitive performance with lower source-workflow latency (Table 1). It wins five of six performance aggregates against native source workflows and is 2.4–17.7× faster. Mean T2SQL EX is 67.1 versus APEX-SQL's 68.5, at 38% lower cost and 3.6× lower latency. Appendices E.1 and E give Pareto plots and all 11 datasets.

Across 16 dataset–metric tuples, deployed AdaSkill lift falls with output F ($r = -0.83$, $p = 8.3 \times 10^{-5}$) and response-free trace F ($r = -0.79$, $p < 0.001$). This is a system-level pattern, not independent mechanism identification: AdaSkill itself consumes F. The following controls isolate structural transfer.

4.3 RQ1: Why Can the Same Pipeline Help and Harm?

Scorer reversal. Figure 4 visualizes the scorer contrast with 60 displayed CE Base Agent pairs, rather than the full diagnostic budget. Canonical estimator-class F is 0.10 for MSA and 1.03 for MRE. Shared-run pairs are not independent observations.

Estimator identity is not outcome-sufficient: $R_{\pi,B}$ in Eq. (3) retains within-method execution effects. On Textbook, adding response-free configuration—covariates, calls, hyperparameters, and tool paths—gives $F = 0.18$ and 1.09. Removing both final responses and selected-method fields from traces gives $F^{\text{trace}}_{\text{MSA}} = 0.55$ and $F^{\text{trace}}_{\text{MRE}} = 1.12$. Thus richer representations preserve the reversal without assuming a fixed coarse kernel (Appendix Table 5).

The matched treatment reverses on the same instances: Full minus Discard is $+20.0$ normalized points for MSA and -10.3 for MRE on Textbook; on Real, $+25.0$ and -2.1 (Table 7). Source, capabilities, and instances remain fixed. Structural value is metric-conditioned even within one task.

4.4 RQ2: When Does Pipeline Transfer Help?

Atomic decomposition. Figure 5 measures the fraction of Final AdaSkill's Base-to-Final gain recovered by each configuration, computed per dataset before metric aggregation. Tools and knowledge

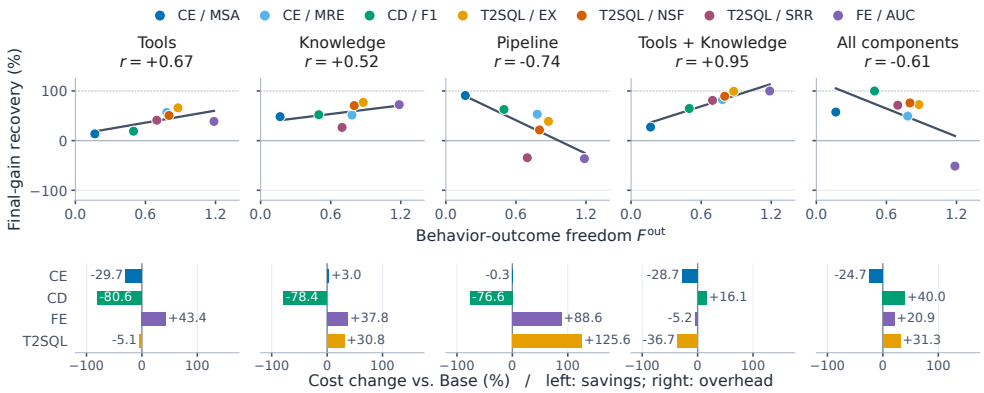


Figure 5: **Capability augmentation and structural constraint have opposite profiles.** Top: fraction of AdaSkill's gain over Base recovered by each configuration. Tools+Knowledge is Discard; All Components is Full. Bottom: cost change relative to Base; negative means savings. Single-component costs follow Table 1; Full/Discard retain the original ablation results.

rise with F ($r = 0.67, 0.52$); their combination strengthens the trend ($r = 0.95$). Pipeline alone reverses it ($r = -0.74$), as does the complete set ($r = -0.61$). Across 15 dataset–metric Pipeline-only treatments versus Base Agent, normalized lift also falls with F (Pearson $r = -0.623$, $p = 0.013$; Spearman $\rho = -0.568$, $p = 0.027$).

Capability-matched interventions. Holding tools, knowledge, evaluator, and execution budget fixed, define $\delta_{\text{pipe}} = u_{\text{full}} - u_{\text{discard}}$; figures use its headroom-normalized counterpart τ_{pipe}. Across 16 interventions, workflow-regime fixed effects give a native-scale slope of -9.18 utility points per unit F (dataset-clustered 95% CI $[-13.41, -4.94]$, $p < 0.001$). Native-scale $r = -0.798$ ($p = 0.0002$); sign ROC AUC is 1.000 (Figure 6). Requiring at least three identifiable questions retains 13 interventions: slope -9.72 (CI $[-14.00, -5.45]$, $p < 0.001$), AUC 1.000. Family-excluded slopes remain negative (-5.58 to -9.87). Appendix C.2 reports every treatment and inference specification.

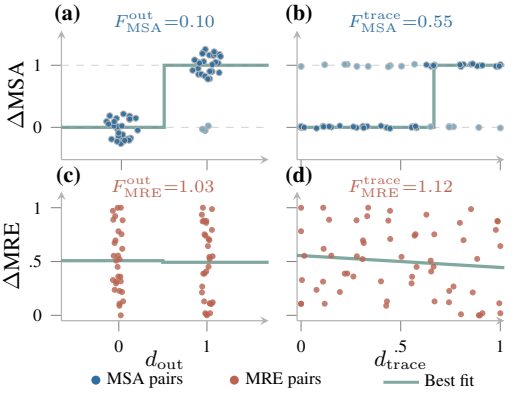


Figure 4: F **quantifies behavior-outcome coupling.** Each point pairs two Base Agent runs. Changing only the scorer separates MSA from MRE in output and response-free trace spaces.

Baseline utility does not explain treatment effects ($r = 0.033$, $p = 0.904$). Unsigned distance dependence recovers association but not its orientation; F distinguishes concordance from anti-concordance, while executed interventions measure candidate direction. Dataset–metric pairs are the estimands; clustering keeps metrics sharing a dataset in one dependence block.

4.5 RQ3: Workflow-Excluded Construction

Frozen routing. Each fold excludes the target workflow regime, calibrates t on the other regimes' aggregated modules, and freezes it before construction. Seven module inputs are CE MSA 0.167, CE MRE 0.782, CD F1 0.602, T2SQL EX 0.686, NSF 0.816, unidentifiable SRR, and FE AUC 1.037. CE/CD aggregate diagnostic datasets; T2SQL/FE use BIRD/Taobao. Six modules qualify; SRR defaults to Discard. These are not the 16 RQ2 coordinates. CD/FE execute the selected candidate; CE/T2SQL directly execute one jointly composed skill.

Mean normalized lift is 31.09 versus 22.11 for Always Full, 21.90 for Always Discard, and 33.71 for full-data AdaSkill (Figure 7). These are seven equally weighted modules: CE/CD average their diagnostic datasets; T2SQL/FE use BIRD/Taobao. LOFO preserves 92.2% of the full-data result.

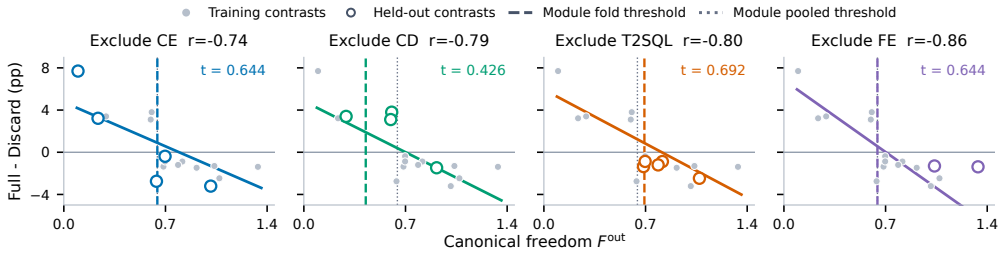


Figure 6: **Behavior-outcome geometry orders the value of pipeline transfer.** Lines summarize 16 dataset-level contrasts; open points are excluded regimes. Thresholds are calibrated separately on aggregated modules, not these plotted points (Appendix A.10). Native-scale $r = -0.798$; descriptive ROC AUC $= 1.000$.

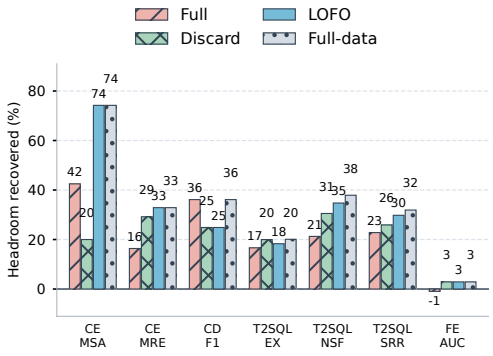


Figure 7: **Workflow-excluded routing.** Headroom recovery (rounded %) for seven modules; native results: Appendix Table 8.

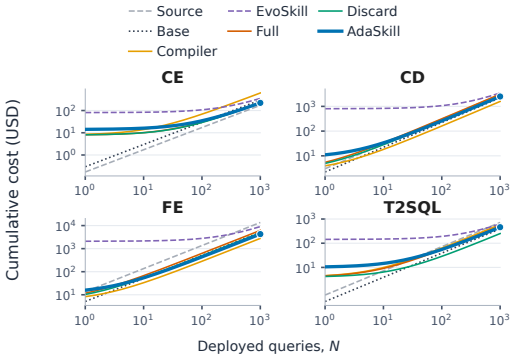


Figure 8: **Cumulative cost projections.** Construction plus N executions, under the accounting in Appendix A.10; both axes are logarithmic.

With skills and thresholds fixed, LOFO exceeds both fixed policies in 97.8% of paired evaluation-bootstrap samples (Appendix A.10). Gains reflect routing and synthesis jointly, not routing alone.

Deployment accounting. Figure 8 projects cumulative cost as $C_0 + Nc$, combining a one-time construction charge with task-average execution costs. Compiler, Full, and Discard share a conversion charge; AdaSkill adds a fixed \$6.12 diagnostic allowance per task family, while EvoSkill includes iterative construction. Source and Base have no new construction charge. AdaSkill's lower execution cost amortizes its initial overhead against several alternatives; rankings depend on task and deployment volume. Appendix A.10 specifies the accounting parameters.

4.6 Sensitivity and Uncertainty Analysis

Diagnostic budget. Figure 9 varies questions $M = 1$–10 and runs $N = 3$–12. MSA stabilizes at $N \geq 5$, MRE stays on the discard side, and $M > 6$ brings marginal movement. CE uses $M = N = 6$: at \$0.17/run, \$6.12 is $2.8\times$ cheaper than $M = N = 10$. T2SQL uses $M = 10, N = 6$.

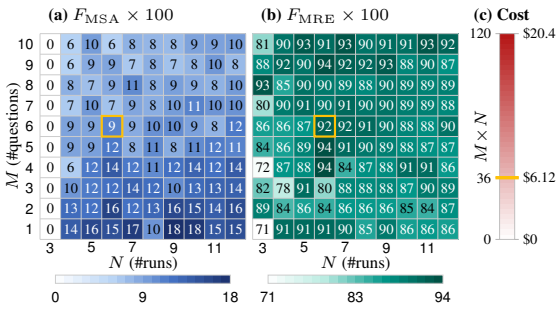


Figure 9: **Budget sensitivity of F_{out}.** The gold operating point ($M{=}N{=}6$) costs \$6.12; estimates stabilize beyond $N{\geq}5$.

Sampling and backbone. Independent runs recover the planner's CE scorer ordering with sufficient budget. At $M = N = 20$, planner/no-planner trace F is $0.55/0.53$ for MSA and $0.85/0.85$ for MRE. Thus the planner improves coverage speed; the ordering persists under a changed behavior law. GPT-5.1 also preserves the system-level output/trace relationship ($r = -0.71, -0.79$; Appendix C.5).

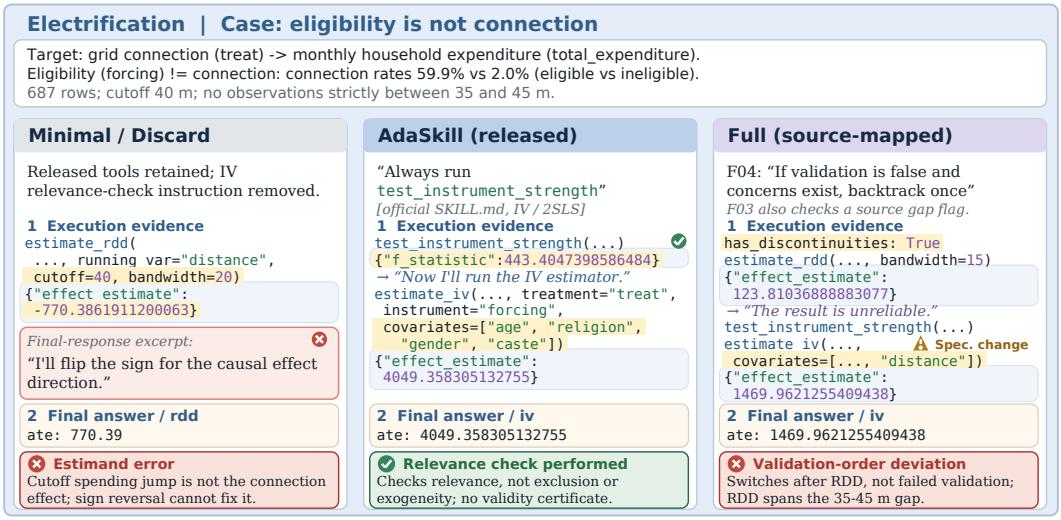


Figure 10: **Eligibility is not treatment.** Exploratory executions with identical data, tools, and Sonnet 4.6 expose estimand, specification, and validation-order differences. Reflowed excerpts and audit icons are not guarantees of correctness or IV validity (Appendix F.1).

Uncertainty. AdaSkill reduces mean CE MRE by 5.6 points versus Base (paired 95% CI $[1.0, 10.5]$). Intervals excluding zero include Textbook MSA, Synthetic MRE, and BIRD EX versus Base, and Spider EX versus Compiler; Spider EX versus Base remains unresolved. Appendix C.4 reports question-level and paired gain intervals for directly executed skills.

Execution case: eligibility is not treatment. Figure 10 presents a new exploratory execution on an already inspected CE case, not another held-out result. The released skill retains IV relevance checking; Minimal mistakes a sign-flipped eligibility jump for the effect of connection under incomplete compliance. Full reaches IV after an RDD detour but changes covariates. The case reveals estimand error and within-method variation, not clause-level causality (replay and construction: Appendix F.1).

Matched construction. AdaSkill uses the released CE skill verbatim. Minimal deletes pipeline clauses but retains separable knowledge, tools, and the output interface. Full adds source-mapped CAIS ordering, validation, and backtracking within supported tools; it is not a native MAS execution. All arms share the CSV, tool code, model, dependencies, and per-run budget. Tool-provided advice remains available even to Minimal. Each arm completes one execution without retries, seed search, or result-dependent revision; no Freedom value or threshold is re-estimated.

Execution-level interpretation. Eligibility is not connection: rates are 59.9% versus 2.0% for eligible versus ineligible households. A sign flip cannot correct incomplete compliance. AdaSkill and Full both report IV, but Full adds distance; numerical replay reproduces 4049.36 versus 1469.96. The shared method label therefore hides specification changes, illustrating the need for $R_{\pi,B}$. One execution per condition establishes neither clause-level causality nor IV validity; relevance alone does not establish exclusion or exogeneity.

5 Conclusion

AdaSkill treats workflow distillation as selective transfer: preserve validated capabilities, remove orchestration, and condition pipeline inheritance on behavior–outcome geometry. Candidate-conditional theory and an explicit representation remainder explain the mechanism; module-level calibration supplies the executable threshold. Scorer controls and matched interventions isolate structural instability, while workflow-excluded construction tests directly executed skills. Across 11 datasets, selective inheritance combines strong performance with lower recurring overhead. The execution case further shows why retaining a method name need not preserve its estimand, specification, or validation order.

AI USE STATEMENT

Generative AI tools were used to adapt the manuscript to the official ICLR 2027 LaTeX template, retrieve related literature, draft and revise text, assist with mathematical proof development and verification, reorganize the layout, and check visual formatting. For the exploratory execution case in Figure 10, an AI coding assistant also helped construct the source-mapped controls, launch model executions, audit logs, independently replay point estimates, and prepare the visualization. The case uses an existing dataset and released estimator implementations; its numerical results are execution outputs, not illustrative or generated measurements. The authors are responsible for reviewing all AI-assisted changes and for the final content, including text, code, and artifacts produced with generative AI.

REPRODUCIBILITY STATEMENT

The main paper specifies the estimands, evaluation protocol, controlled interventions, and deployment rule. The appendix provides complete proofs, source-workflow qualification details, implementation and validation contracts, per-dataset results, robustness analyses, and execution-cost boundaries. All routing thresholds are selected only on development regimes and then frozen; final metrics come from direct execution of the resulting skills. The source package records behavioral representations, diagnostic budgets, conversion prompts, and clause-audit contracts, with aggregate inputs and scripts for replaying threshold calibration and policy summaries. Execution-cost comparisons exclude construction and diagnosis.

REFERENCES

Sawal Acharya, Terry Jingchen Zhang, Andrew Kim, Anahita Haghighat, Sun Xianlin, Pepijn Cobben, Rahul Babu Shrestha, Maximilian Mordig, Jacob T. Emmerson, Furkan Danisman, Yuen Chen, Clijo Jose, Andrei Ioan Muresanu, Justin Cui, Jiarui Liu, Yahang Qi, Punya Syon Pandey, Yinya Huang, Bernhard Schölkopf, Mrinmaya Sachan, and Zhijing Jin. Causcibench: A comprehensive benchmark on end-to-end causal inference for scientific research, 2026. URL https://openreview.net/forum?id=uQzPkWvTyo.

Alibaba Group. Ad SVR prediction data on Taobao.com. https://tianchi.aliyun.com/dataset/147588, 2018.

Aayush Aluru, Myra Malik, Samarth Patankar, Spencer Kim, Kevin Zhu, Sean O'Brien, and Vasu Sharma. Smagdi: Socratic multi agent interaction graph distillation for efficient high accuracy reasoning. *arXiv preprint arXiv:2511.05528*, 2025.

Salaheddin Alzubi, Noah Provenzano, Jaydon Bingham, Weiyuan Chen, and Tu Vu. Evoskill: Automated skill discovery for multi-agent systems. *arXiv preprint arXiv:2603.02766*, 2026.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: 10.3982/ECTA15732.

Bowen Cao, Weibin Liao, Yushi Sun, Dong Fang, Haitao Li, and Wai Lam. Apex-sql: Talking to the data via agentic exploration for text-to-sql. *arXiv preprint arXiv:2602.16720*, 2026.

Justin Chih-Yao Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. MAGDi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. *arXiv preprint arXiv:2402.01620*, 2024. URL https://arxiv.org/abs/2402.01620.

Zhongxin Guo, Danrui Qi, Hanwen Gu, Peng Cheng, and Yongqiang Xiong. SKILL-DISCO: Distilling and compiling agent traces into reusable procedural skills. *arXiv preprint arXiv:2606.26669*, 2026.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin SU, ZHAOQING SUO, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, Qian Liu, Sida Wang, and Tao Yu. Spider 2.0: Evaluating language models on real-world enterprise text-to-SQL workflows. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XmProj9cPs.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36:42330–42357, 2023.

Xiaoxiao Li. When single-agent with skills replace multi-agent systems and when they fail. *arXiv preprint arXiv:2601.04748*, 2026.

Junlin Liu, Jiangwang Chen, Zixin Song, Shuaiyu Zhou, Chunji Lv, Hank Wu, Kailin Jiang, Jinyang Wu, Bohan Yu, and Chenxi Zhou. From proprietary to open-source: Bridging the distribution gap via multi-agent protocol distillation in agentic search. *arXiv preprint arXiv:2607.24280*, 2026.

Yinyi Luo, Yiqiao Jin, Weichen Yu, Mengqi Zhang, Srijan Kumar, Xiaoxiao Li, Weijie Xu, Xin Chen, and Jindong Wang. Agentark: Distilling multi-agent intelligence into a single llm agent. *arXiv preprint arXiv:2602.03955*, 2026.

Yuchen Ma, Yue Huang, Han Bao, Haomin Zhuang, Swadheen Shukla, Michel Galley, Xiangliang Zhang, and Stefan Feuerriegel. Skillgen: Verified inference-time agent skill synthesis. *arXiv preprint arXiv:2605.10999*, 2026.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2_Part_1):209–220, 1967.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17:1103–1204, 2016. URL https://api.semanticscholar.org/CorpusID:272555168.

Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Xiaoxi Jiang, and Guanjun Jiang. Trace2skill: Distill trajectory-local lessons into transferable agent skills, 2026. URL https://arxiv.org/abs/2603.25158.

OpenAI. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/, 2024.

Kun Ouyang, Haoyu Wang, and Dong Fang. Fela: A multi-agent evolutionary system for feature engineering of industrial event log data. *arXiv preprint arXiv:2510.25223*, 2025.

Jiahao Qiu, Xinzhe Juan, Yimin Wang, Ling Yang, Xuan Qi, Tongcheng Zhang, Jiacheng Guo, Yifu Lu, Zixin Yao, Hongru Wang, Shilong Liu, Xun Jiang, Liu Leqi, and Mengdi Wang. AgentDistill: Training-free agent distillation with generalizable MCP boxes. *arXiv preprint arXiv:2506.14728*, 2025. URL https://arxiv.org/abs/2506.14728.

R. Quinlan. Auto MPG [Dataset]. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

ChengAo Shen, Zhengzhang Chen, Dongsheng Luo, Dongkuan Xu, Haifeng Chen, and Jingchao Ni. Exploring multi-modal data with tool-augmented llm agents for precise causal discovery. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 636–660, 2025.

David J Spiegelhalter. Learning in probabilistic expert systems. *Bayesian Statistics*, 4:447–465, 1992.

Alex Teboul. Diabetes health indicators dataset, 2022. URL https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

Vishal Verma, Sawal Acharya, Samuel Simko, Devansh Bhardwaj, Anahita Haghighat, Dominik Janzing, Mrinmaya Sachan, Zhijing Jin, and Yongjin Yang. Causal ai scientist: Facilitating causal data science with large language models. In *NeurIPS 2025 AI for Science Workshop*, 2025.

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. *arXiv preprint arXiv:2305.16291*, 2023.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning, 2026. URL https://arxiv.org/abs/2602.08234.

Binyan Xu, Xilin Dai, and Kehuan Zhang. Contextual agentic memory is a memo, not true memory. *arXiv preprint arXiv:2604.27707*, 2026a.

Binyan Xu, Haitao Li, and Kehuan Zhang. Llm agents are latent context managers: Eliciting self-managed context via a proprioceptive dashboard. *arXiv preprint arXiv:2606.30005*, 2026b.

Jiawei Xu, Arief Koesdwiady, Sisong Bei, Yan Han, Baixiang Huang, Dakuo Wang, Yutong Chen, Zheshen Wang, Peihao Wang, Pan Li, et al. Rethinking the value of multi-agent workflow: A strong single agent baseline. *arXiv preprint arXiv:2601.12307*, 2026c.

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

Yutao Yang, Junsong Li, Qianjun Pan, Bihao Zhan, Yuxuan Cai, Lin Du, Jie Zhou, Kai Chen, Qin Chen, Xin Li, Bo Zhang, and Liang He. Autoskill: Experience-driven lifelong learning via skill self-evolution, 2026b. URL https://arxiv.org/abs/2603.01145.

John Seon Keun Yi, Aaron Mueller, and Dokyun Lee. Latent agents: A post-training procedure for internalized multi-agent debate. *arXiv preprint arXiv:2604.24881*, 2026.

Tik Yu Yim, Wenting Tan, Sum Yee Chan, Tak-Wah Lam, and Siu Ming Yiu. Asda: Automated skill distillation and adaptation for financial reasoning, 2026. URL https://arxiv.org/abs/2603.16112.

Boqin Yuan, Renchu Song, Yue Su, Sen Yang, and Jing Qin. Clawtrace: Cost-aware tracing for llm agent skill distillation. *arXiv preprint arXiv:2604.23853*, 2026.

Huichi Zhou, Siyuan Guo, Anjie Liu, Zhongwei Yu, Ziqin Gong, Bowen Zhao, Zhixun Chen, Menglong Zhang, Yihang Chen, Jinsong Li, Runyu Yang, Qiangbin Liu, Xinlei Yu, Jianmin Zhou, Na Wang, Chunyang Sun, and Jun Wang. Memento-skills: Let agents design agents, 2026. URL https://arxiv.org/abs/2603.18743.

A Theoretical Scope and Proofs

This appendix separates mechanism from decision. It proves a structural decomposition with an explicit representation-kernel remainder, establishes the geometric and anchor-transfer roles of F, and derives the question-mean sufficient threshold available only under candidate qualification bounds. It then proves the empirical threshold-policy bridge actually used by AdaSkill. Sign, scale, support, and multi-module boundaries delimit each claim.

A.1 Formal Setup and Statements

Let X denote the task instance, H the complete evaluator-facing execution state, $B = \phi(H)$ a diagnostic representation, and $Y \in \mathbb{R}$ a higher-is-better score. At any chosen B, write the induced baseline and actual candidate laws as

$$P_j := \nu\mu_j K_j, \qquad j \in \{0, \pi\}, \tag{7}$$

where $\mu_j(db \mid x)$ is the behavior law and $K_j(dy \mid x, b)$ the conditional score law. Even when the evaluator is fixed conditional on H, coarsening can make $K_\pi(\cdot \mid X, B) \neq K_0(\cdot \mid X, B)$ by changing the mixture of execution states inside one B cell. To isolate behavioral transport, define the virtual fixed-kernel intervention $\widetilde{P}_\pi := \nu\mu_\pi K_0$. Assume $\mu_\pi(\cdot \mid x) \ll \mu_0(\cdot \mid x)$, let $w_{\pi,B} = d\mu_\pi/d\mu_0$ admit a jointly measurable $L^2(P_0)$ version, and assume $Y \in L^2(P_0)$ and $\mathbb{E}_{P_\pi}|Y| < \infty$. Define

$$m_0(x,b) = \mathbb{E}_{K_0}[Y \mid X = x, B = b], \qquad m_\pi(x,b) = \mathbb{E}_{K_\pi}[Y \mid X = x, B = b], \tag{8}$$

$$\mu_0(x) = \mathbb{E}_0[Y \mid X = x], \qquad G_B^2 = \mathbb{E}_0\{m_0(X,B) - \mu_0(X)\}^2, \tag{9}$$

$$S_{\pi,B}^2 = \mathbb{E}_0(w_{\pi,B} - 1)^2 = \chi^2_{\pi,B}, \qquad a_{\pi,B} = \frac{\mathbb{E}_0[(w_{\pi,B} - 1)(m_0 - \mu_0)]}{S_{\pi,B} G_B}, \tag{10}$$

with $a_{\pi,B} = 0$ when $S_{\pi,B} G_B = 0$.

Theorem A.1 (Structural-Lift Decomposition)**.** *Under the conditions above,*

$$T_B(\pi) := \mathbb{E}_{\widetilde{P}_\pi} Y - \mathbb{E}_{P_0} Y = a_{\pi,B} S_{\pi,B} G_B, \qquad |T_B(\pi)| \le S_{\pi,B} G_B. \tag{11}$$

For conditionally independent $B, B' \sim \mu_0(\cdot \mid X)$,

$$G_B^2 = \frac{1}{2}\mathbb{E}\{m_0(X,B) - m_0(X,B')\}^2. \tag{12}$$

The actual lift obeys the exact representation-aware decomposition

$$\mathrm{Lift}(\pi) = T_B(\pi) + R_{\pi,B}, \qquad R_{\pi,B} := \mathbb{E}_{\nu\mu_\pi}[m_\pi(X,B) - m_0(X,B)]. \tag{13}$$

For the pair variables U and V in Eq. (1), let $\widetilde{R}_U$ and $\widetilde{R}_V$ be their centered, variance-one population mid-ranks and define

$$\Gamma_{\mathrm{rank}} := \mathrm{Var}\Big(\mathbb{E}[\widetilde{R}_V \mid U]\Big). \tag{14}$$

Proposition A.2 (Freedom Probes Geometric Determinacy)**.** *When both ranks are non-degenerate,*

$$0 \le \rho_S(U,V)^2 \le \Gamma_{\mathrm{rank}} \le 1, \qquad 1 - \Gamma_{\mathrm{rank}} \le 2F - F^2. \tag{15}$$

The first bound is tight if and only if $\mathbb{E}[\widetilde{R}_V \mid U] = \rho_S(U,V)\widetilde{R}_U$ almost surely.

Let $Z \sim P_0$, let A and C be the standardized population mid-ranks of U and V, and fix a high-score anchor event $\mathcal{H}$ with $h = P_0(\mathcal{H}) > 0$ and $Q = P_0(\cdot \mid \mathcal{H})$. For $\widetilde{P}_\pi \ll P_0$, define

$$\Delta_\pi := \mathbb{E}_{P_0 \otimes Q} A(Z, Z^\star) - \mathbb{E}_{\widetilde{P}_\pi \otimes Q} A(Z, Z^\star), \qquad \Lambda_\pi := \mathbb{E}_{P_0 \otimes Q} C(Z, Z^\star) - \mathbb{E}_{\widetilde{P}_\pi \otimes Q} C(Z, Z^\star). \tag{16}$$

Theorem A.3 (Signed Anchor-Rank Transfer)**.** *If $\chi^2_{\pi,B} = \mathbb{E}_0(w_{\pi,B} - 1)^2 < \infty$, then for either $\kappa \in \{-1, +1\}$,*

$$|\Lambda_\pi - \kappa\Delta_\pi| \le \sqrt{\frac{2\chi^2_{\pi,B}\{1 - \kappa\rho_S(U,V)\}}{h}}. \tag{17}$$

The theorem above deliberately applies to $\widetilde{P}_\pi$ and therefore uses behavior divergence $\chi^2_{\pi,B}$. Applied directly to an actual candidate law, the same algebra requires the joint-run ratio dP_π/dP_0 and the joint divergence $\chi^2(P_\pi \| P_0)$; in that alternative formulation kernel shift is already included and must not be counted again as $R_{\pi,B}$.

A.2 Proof of the Structural-Lift Decomposition

Write $g_B(X,B) = m_0(X,B) - \mu_0(X)$ and $h_{\pi,B}(X,B) = w_{\pi,B}(X,B) - 1$. Because $w_{\pi,B}$ is a conditional density ratio,

$$\mathbb{E}_0[h_{\pi,B} \mid X] = 0. \tag{18}$$

The common conditional outcome kernel gives

$$T_B(\pi) = \mathbb{E}_0[w_{\pi,B}(X,B)m_0(X,B)] - \mathbb{E}_0 m_0(X,B) \tag{19}$$

$$= \mathbb{E}_0[h_{\pi,B}m_0] = \mathbb{E}_0[h_{\pi,B}g_B]. \tag{20}$$

The two factors in the last inner product have $L^2(P_0)$ norms $S_{\pi,B}$ and G_B. Their normalized inner product is $a_{\pi,B}$, proving Eq. (11). Cauchy–Schwarz gives $|T_B(\pi)| \le S_{\pi,B}G_B$. When $S_{\pi,B}G_B > 0$, equality holds if and only if $h_{\pi,B} = cg_B$ P_0-almost surely for a constant c compatible with $w_{\pi,B} \ge 0$.

For conditionally independent $B, B' \sim \mu_0(\cdot \mid X)$,

$$\mathbb{E}[\{m_0(X,B) - m_0(X,B')\}^2 \mid X] = 2\,\mathrm{Var}(m_0(X,B) \mid X). \tag{21}$$

Averaging over X proves Eq. (12). Adding and subtracting $\mathbb{E}_{\nu\mu_\pi} m_0$ proves Eq. (13). Thus an evaluator-sufficient H has zero remainder under a fixed harness, while a coarse $B = \phi(H)$ remains valid through the displayed remainder rather than through an invariance assertion. If $R_{\pi,B} \ge -\epsilon_K$, any lower bound on T_B transfers to actual lift after subtracting ϵ_K.

If μ_π has off-support mass ω and its normalized on-support part has lower bound L_o, bounded utility with range R only yields

$$\mathrm{Lift}(\pi) \ge (1-\omega)L_o - \omega R. \tag{22}$$

The on-support term must be scaled; merely subtracting the support-violation rate from the original dominated-law bound is not valid. Behavioral overlap is a condition of the mechanism bound, not an implemented statistical test. The converter checks tool interfaces and clause validity; finite diagnostic runs do not certify absolute continuity or bound off-support mass. Routing therefore relies on development interventions, not an overlap certificate.

A.3 Proof of Geometric Determinacy

Let $X_U = \widetilde{R}_U$ and $X_V = \widetilde{R}_V$. Both have mean zero and variance one. Since X_U is measurable with respect to U,

$$\rho_S(U,V) = \mathbb{E}(X_U X_V) = \mathbb{E}[X_U \mathbb{E}(X_V \mid U)]\,. \tag{23}$$

Cauchy–Schwarz yields

$$\rho_S(U,V)^2 \le \mathbb{E}X_U^2\, \mathbb{E}\big[\{\mathbb{E}(X_V \mid U)\}^2\big] = \Gamma_{\mathrm{rank}}. \tag{24}$$

Because $\mathbb{E}\{\mathbb{E}(X_V \mid U)\} = 0$, the second moment above equals $\mathrm{Var}\{\mathbb{E}(X_V \mid U)\}$. The law of total variance gives $0 \le \Gamma_{\mathrm{rank}} \le \mathrm{Var}(X_V) = 1$. Substituting $\rho_S = 1 - F$ proves Eq. (15). Equality holds if and only if $\mathbb{E}(X_V \mid U) = cX_U$ almost surely for a constant c, equivalently when the unstandardized conditional rank expectation is affine in R_U.

The canonical question-conditioned mean inherits a corresponding guarantee. Applying the proposition to each identifiable question and then Jensen's inequality gives

$$\mathbb{E}_q\Gamma_q \ge \mathbb{E}_q(1-F_q)^2 \ge (1 - \mathbb{E}_q F_q)^2. \tag{25}$$

This is the same mean aggregate consumed by the converter and used by the structural analysis.

This result separates signed monotone coupling from optimal L^2 prediction of outcome-distance rank from behavioral distance. Γ_{rank} is the corresponding correlation ratio, while Spearman correlation uses the single predictor $\widetilde{R}_U$. Consequently, $F \approx 1$ indicates weak monotone coupling, but does not exclude a nonlinear dependence. When $F > 1$, the monotone component is reversed; a value near two can be highly predictive in the anti-concordant direction. Writing $D = |\rho_S|$ and $O = \mathrm{sign}(\rho_S) \in \{-1, 0, +1\}$ makes the distinction explicit: D measures unsigned monotone coupling strength, while O gives positive, absent, or negative orientation. The next theorem uses the two nonzero orientations explicitly.

A.4 Proof of Signed Anchor-Rank Transfer

Recall that A and C are standardized population mid-ranks of $U = d(\phi(Z), \phi(Z'))$ and $V = |s(Z) - s(Z')|$, respectively. Standardization gives $\mathbb{E}A = \mathbb{E}C = 0$, $\mathbb{E}A^2 = \mathbb{E}C^2 = 1$, and $\mathbb{E}(AC) = \rho_S(U, V) = 1 - F$. Therefore, for either $\kappa \in \{-1, +1\}$,

$$\mathbb{E}(A - \kappa C)^2 = 2\{1 - \kappa\rho_S(U, V)\}. \tag{26}$$

For $\kappa = +1$ this is $2F$; for $\kappa = -1$ it is $2(2 - F)$. The identity includes discrete variables because the population definition uses mid-ranks.

Proof of Theorem A.3. Fix $\kappa \in \{-1, +1\}$ and let $E_\kappa = A - \kappa C$. Let $\mathcal{H}$ be the fixed anchor event, $h = P_0(\mathcal{H}) > 0$, and $Z^\star \sim Q = P_0(\cdot \mid \mathcal{H})$. Conditioning the second baseline run on the anchor event and using Eq. (26) yields

$$\mathbb{E}_{\mathbb{P}_0 \otimes Q} E_\kappa^2 = \frac{\mathbb{E}_{\mathbb{P}_0 \otimes \mathbb{P}_0}[E_\kappa^2 \mathbf{1}\{Z^\star \in \mathcal{H}\}]}{h} \le \frac{2\{1 - \kappa\rho_S(U, V)\}}{h}. \tag{27}$$

Using $E_\kappa = A - \kappa C$ and $d\widetilde{\mathbb{P}}_\pi = w_{\pi,B}\, d\mathbb{P}_0$ in the first coordinate,

$$\kappa\Lambda_\pi - \Delta_\pi = \mathbb{E}_{\mathbb{P}_0 \otimes Q}[(w_{\pi,B} - 1)E_\kappa]. \tag{28}$$

Cauchy–Schwarz and Eq. (27) give

$$|\kappa\Lambda_\pi - \Delta_\pi| \le \sqrt{\mathbb{E}_{\mathbb{P}_0}(w_{\pi,B} - 1)^2}\sqrt{\mathbb{E}_{\mathbb{P}_0 \otimes Q} E_\kappa^2} \le \sqrt{\frac{2\chi^2_{\pi,B}\{1 - \kappa\rho_S(U, V)\}}{h}}. \tag{29}$$

Using $|\kappa x| = |x|$ proves Eq. (17). □

A.5 Calibration to Continuous Metric Lift

Signed Anchor-Rank Transfer is exact whenever both distance-rank variables are non-degenerate because F and Λ_π live on the same ordinal score-distance scale. The following corollary quantifies the only additional step required to recover mean lift on a continuous metric's original scale.

Corollary A.4 (Calibrated Continuous-Score Transfer)**.** *Under Theorem A.3, let*

$$g_{\mathcal{H}}(y) := \mathbb{E}_{Z^\star \sim Q}[C(Z, Z^\star) \mid Y = y] = a_{\mathcal{H}} - b_{\mathcal{H}} y + \varepsilon_{\mathcal{H}}(y), \qquad b_{\mathcal{H}} > 0, \tag{30}$$

where $a_{\mathcal{H}} - b_{\mathcal{H}} y$ is the $L^2(P_0)$ affine projection and $\sigma^2_{\mathcal{H}} = \mathbb{E}_0 \varepsilon_{\mathcal{H}}(Y)^2$. Then, for either $\kappa \in \{-1, +1\}$,

$$\left| T_B(\pi) - \frac{\kappa\Delta_\pi}{b_{\mathcal{H}}} \right| \le \frac{1}{b_{\mathcal{H}}} \left[\sqrt{\frac{2\chi^2_{\pi,B}\{1 - \kappa\rho_S(U, V)\}}{h}} + \sqrt{\chi^2_{\pi,B}}\, \sigma_{\mathcal{H}} \right]. \tag{31}$$

Proof. Because C depends on the score pair only through $V = |Y - Y^\star|$,

$$\Lambda_\pi = \mathbb{E}_{P_0} g_{\mathcal{H}}(Y) - \mathbb{E}_{\widetilde{P}_\pi} g_{\mathcal{H}}(Y) \tag{32}$$

$$= b_{\mathcal{H}} T_B(\pi) - \mathbb{E}_{P_0}[(w_{\pi,B} - 1)\varepsilon_{\mathcal{H}}(Y)]. \tag{33}$$

Hence

$$|\Lambda_\pi - b_{\mathcal{H}} T_B(\pi)| \le \sqrt{\chi^2_{\pi,B}}\, \sigma_{\mathcal{H}} \tag{34}$$

by Cauchy–Schwarz. Combining this inequality with $|\Lambda_\pi - \kappa\Delta_\pi| \le \sqrt{2\chi^2_{\pi,B}\{1 - \kappa\rho_S(U, V)\}/h}$ from Theorem A.3, then dividing by $b_{\mathcal{H}}$, proves the result. □

The corollary applies to bounded scores including MRE utility, F1, AUC, NSF, and binary SRR after their fixed higher-is-better orientation. It does not assume perfect-score observations. A score-threshold or quantile rule may be fixed from the baseline law before intervention evaluation; $b_{\mathcal{H}}$ and $\sigma_{\mathcal{H}}$ are then baseline-estimable. A small residual makes rank transfer informative on the native metric scale, while a large residual correctly exposes that an ordinal diagnostic cannot determine raw-scale gain. If $b_{\mathcal{H}} \le 0$, the chosen anchors do not define a positively calibrated metric direction and this metric-scale corollary abstains; the ordinal theorem remains valid.

A.6 Exact Binary Specialization

Let $Y \in \{0,1\}$, $p = P_0\{Y = 1\} \in (0,1)$, and $\mathcal{H} = \{Y = 1\}$. Because the anchor succeeds, $V(Z, Z^\star) = 1 - Y$. The standardized mid-rank C therefore takes one value c_0 when $Y = 1$ and another c_1 when $Y = 0$. Since V is Bernoulli with probability $q = 2p(1-p)$ under the baseline product law,

$$c_1 - c_0 = \frac{1}{\sqrt{q(1-q)}} = \frac{1}{\beta_p}, \qquad \beta_p = \sqrt{2p(1-p)\{1-2p(1-p)\}}. \tag{35}$$

Thus $g_{\mathcal{H}}(Y) = c_1 - Y/\beta_p$, so $b_{\mathcal{H}} = 1/\beta_p$ and $\sigma_{\mathcal{H}} = 0$. Equivalently,

$$\Lambda_\pi = \mathbb{E}_{P_0 \otimes Q} C - \mathbb{E}_{\widetilde{P}_\pi \otimes Q} C = \frac{T_B(\pi)}{\beta_p}. \tag{36}$$

Corollary A.4 therefore gives

$$|T_B(\pi) - \kappa \beta_p \Delta_\pi| \le \beta_p \sqrt{\frac{2\chi^2_{\pi,B}\{1 - \kappa \rho_S(U,V)\}}{p}}. \tag{37}$$

The two orientations yield distinct transfer statements. For $\kappa = +1$, low F gives $T_B(\pi) \approx \beta_p \Delta_\pi$. For $\kappa = -1$, F near two gives $T_B(\pi) \approx -\beta_p \Delta_\pi$, or

$$|T_B(\pi) + \beta_p \Delta_\pi| \le \beta_p \sqrt{\frac{2\chi^2_{\pi,B}(2-F)}{p}}. \tag{38}$$

In particular, if $F > 1$ and

$$\Delta_\pi > \sqrt{\frac{2\chi^2_{\pi,B}(2-F)}{p}}, \tag{39}$$

then the upper side of Eq. (38) is negative and $T_B(\pi) < 0$. Thus strong anti-concordance can certify that successful-anchor contraction is harmful for the behavior-mediated component. Values only slightly above one need not satisfy this certificate; they still provide no positive-orientation guarantee and are assigned to the discard side of AdaSkill's one-sided policy.

A.7 Question-Conditioned Anchor Transfer

For binary question-level metrics, the positive-transfer statement uses orientation $\kappa = +1$. A calibrated operating point $t < 1$ excludes a metric-level statistic with nonpositive orientation; individual question-level estimates above one remain valid and enlarge the positive-transfer error term. Their opposite-sign implication is given separately by Eq. (38).

Corollary A.5 (Question-Conditioned Anchor Transfer)**.** *Let $\mathcal{I}$ be the population set of evaluation questions for which a binary score has non-degenerate behavioral and score-distance ranks and $0 < p_x < 1$. Let $\nu_{\mathcal{I}} = \nu(\cdot \mid \mathcal{I})$. Suppose the binary specialization of Theorem A.3 with $\kappa = +1$ holds conditionally for $\nu_{\mathcal{I}}$-almost every x, with quantities p_x, β_x, F_x, $\Delta_{\pi,x}$, and finite $\chi^2_{\pi,B,x}$, and assume $\mathbb{E}_{\nu_{\mathcal{I}}}[\beta_x^2 \chi^2_{\pi,B,x}/p_x] < \infty$. Then*

$$|\mathbb{E}_{x \sim \nu_{\mathcal{I}}} T_{B,x}(\pi) - \mathbb{E}_{x \sim \nu_{\mathcal{I}}}[\beta_x \Delta_{\pi,x}]| \le \mathbb{E}_{x \sim \nu_{\mathcal{I}}}\left[\beta_x \sqrt{\frac{2\chi^2_{\pi,B,x} F_x}{p_x}}\right] \tag{40}$$

$$\le \sqrt{2\mathbb{E}_{x \sim \nu_{\mathcal{I}}}[F_x]\mathbb{E}_{x \sim \nu_{\mathcal{I}}}\left[\frac{\beta_x^2 \chi^2_{\pi,B,x}}{p_x}\right]}. \tag{41}$$

Proof. Apply the binary specialization in Eq. (37) conditionally on x, take expectations, and use $|\mathbb{E}R| \le \mathbb{E}|R|$ for the first inequality. The second follows from Cauchy–Schwarz applied to $\sqrt{F_x}$ and $\beta_x \sqrt{\chi^2_{\pi,B,x}/p_x}$. □

The corollary bounds mean behavior-mediated lift over population-identifiable questions. Define

$$M_{\pi,I} := \mathbb{E}_I[\beta_x \Delta_{\pi,x}], \qquad C_{\pi,I} := \mathbb{E}_I[\beta_x^2 \chi^2_{\pi,B,x}/p_x], \qquad \bar{F}_I := \mathbb{E}_I F_x. \tag{42}$$

If $M_{\pi,I} \geq m > e_K \geq 0$, $C_{\pi,I} \leq c$ with $c > 0$, and the mean representation remainder is at least $-e_K$, then

$$\mathbb{E}_I \mathrm{Lift}_x(\pi) \geq m - e_K - \sqrt{2c\bar{F}_I}, \qquad \bar{F}_I < \frac{(m-e_K)^2}{2c} \implies \mathbb{E}_I \mathrm{Lift}_x(\pi) > 0. \tag{43}$$

For $c = 0$, the lower bound reduces to $m - e_K$; no division by c is used. Without $m > e_K$, the lower bound remains valid but gives no positive-lift certificate. This is a sufficient F-threshold only for candidates whose m, c, e_K contracts are estimated or otherwise certified; AdaSkill does not use this certificate as its threshold.

Write $\alpha = \nu(\mathcal{I})$ and let $R_{\mathcal{I}}$ denote the first displayed upper bound in Corollary A.5. Since binary lift has absolute value at most one, the full benchmark satisfies

$$|\mathbb{E}_{x\sim\nu} T_{B,x}(\pi) - \alpha \mathbb{E}_{x\sim\nu_{\mathcal{I}}}[\beta_x \Delta_{\pi,x}]| \leq \alpha R_{\mathcal{I}} + 1 - \alpha. \tag{44}$$

In finite samples, observed score-varying questions estimate a sample event $\mathcal{I}_N$, not population $\mathcal{I}$. AdaSkill therefore treats the mean over observed identifiable questions as a diagnostic coordinate and defaults to Discard below its minimum count; it does not claim a worst-case certificate for the unidentifiable stratum.

A.8 From Geometric Ordering to the Operating Threshold

The preceding sufficient threshold is candidate-conditional and is not the implemented calibration rule. Let a development structural unit have pre-action coordinate F, capability-matched potential utilities (U^R, U^D), and $\delta = U^R - U^D$. For $t \in \overline{\mathbb{R}}$ and $\pi_t(F) = \mathbf{1}\{F < t\}$, define

$$V(t) = \mathbb{E}[U^D + \pi_t(F)\delta], \qquad \widehat{V}(t) = \sum_i \omega_i [U_i^D + \pi_t(F_i)\delta_i], \tag{45}$$

where the extended-real endpoints represent the two constant policies, nonnegative weights sum to one, and metrics sharing executions remain in the same dependence cluster.

Proposition A.6 (Empirical threshold policy)**.** *For $t^\star \in \arg\max_t V(t)$ and $\widehat{t} \in \arg\max_t \widehat{V}(t)$,*

$$V(t^\star) - V(\widehat{t}) \leq 2 \sup_t |\widehat{V}(t) - V(t)|. \tag{46}$$

Moreover, if $\tau(f) = \mathbb{E}[\delta \mid F = f]$ is non-increasing and F has no atom at its sign-change boundary, the strict-threshold class contains a Bayes-optimal F-measurable policy. Without the no-atom condition, both open and closed lower thresholds must be allowed.

Proof. Empirical optimality makes $\widehat{V}(t^\star) - \widehat{V}(\widehat{t}) \leq 0$. Adding and subtracting the two empirical values gives

$$V(t^\star) - V(\widehat{t}) \leq |V(t^\star) - \widehat{V}(t^\star)| + |\widehat{V}(\widehat{t}) - V(\widehat{t})|, \tag{47}$$

which proves Eq. (46). Conditional on $F = f$, Retain is weakly preferable when $\tau(f) \geq 0$; monotonicity makes the positive-effect set a lower interval. Its endpoint may be open or closed. The no-atom condition makes the two choices equal in population value; treatment on a zero-effect interval is immaterial. □

The archived implementation minimizes unweighted sign-classification errors over identifiable metric modules, then maximizes the distance to the nearest training coordinate and breaks remaining ties by the smaller threshold. Candidates are adjacent-coordinate midpoints and two constant-policy endpoints. On these data, every fold separates positive and negative module effects; its selected threshold therefore also maximizes Eq. (45) for any positive module weights. This is an observed equivalence, not equivalence of the objectives in general. The guarantee is best-in-class average value, not individual safety; four regimes do not imply a small population regret bound.

For a multi-metric task, global Full and global Discard observe only two corners of a route vector. They do not identify coordinate effects under arbitrary module interactions. Proposition A.6 justifies the calibrated structural-unit ordering, while RQ3 directly executes the skill; no global optimality claim is made without a separability assumption or factorial interventions.

A.9 Basic Properties of the Protocol-Conditioned Statistic

Proposition A.7 (Range, invariance, and degeneracy)**.** *For pair variables $U = d(\phi(Z), \phi(Z'))$ and $V = |s(Z) - s(Z')|$ whose mid-distribution ranks have nonzero variance:*

(i) $F = 1 - \rho_S(U, V) \in [0, 2]$*;*

(ii) F *is invariant to strictly increasing transformations of* U *or* V*, including positive affine rescaling of metric scores;*

(iii) if U *or* V *is constant,* F *is undefined;*

(iv) changing d*,* ϕ*, or the protocol-induced distribution* $\mathbb{P}_0$ *can change* F*.*

Proof. Part (i) follows from $\rho_S \in [-1, 1]$. Part (ii) follows because strictly increasing transformations preserve ranks. In Part (iii), a constant variable has zero rank variance, so the correlation denominator is zero. Part (iv) follows directly because each of d, ϕ, and $\mathbb{P}_0$ enters the joint law of (U, V) in Equation (1). □

When $F > 1$, $\rho_S < 0$: behaviorally distant runs tend to be closer in score rank than behaviorally similar runs. This is anti-concordance, not independence. The unsigned quantity $1 - |\rho_S|$ recognizes predictability near both endpoints but cannot route a contraction because it erases whether the predicted transfer sign is positive or negative. Signed Anchor-Rank Transfer shows that anchor contraction has the opposite leading effect as F approaches two. Our empirical plots therefore place these observations on the monotone discard side. An intervention deliberately oriented against the source contraction defines a different policy.

A.10 Calibration Replay and Evidence Boundaries

The source package includes `plot_data/result_bundle_input.csv`, the aggregate input with author-corrected binary entries. `scripts/audit_revision.py` regenerates candidate thresholds, native and normalized objective values, routes, and module summaries. Hashes identify this audit's inputs, not historical freezing times or executed-skill hashes.

Calibration uses the pre-specified `LOFO scope`: CE MSA averages Textbook/Real; CE MRE averages Textbook/Real/Synthetic; CD F1 averages all four CD datasets; T2SQL uses BIRD; FE uses Taobao. A module qualifies only if all its finite question-identifiability counts reach three. Thus SRR defaults to Discard, leaving six calibration modules. This differs from the 16 dataset–metric contrasts used for mechanism analysis; CE Synthetic MRE enters its module average, not as an independent threshold constraint.

Module	Mean F	Mean $(u^R - u^D)$	LOFO action
CE MSA	.166550	.054585	Retain
CE MRE	.782200	-.021133	Discard
CD F1	.602075	.022075	Discard
T2SQL EX	.685900	-.013600	Retain
T2SQL NSF	.816100	-.008800	Discard
T2SQL SRR	unidentifiable	-.008700	Discard
FE AUC	1.036700	-.013000	Discard

The pooled threshold is .6439875. Excluding CE, CD, T2SQL, and FE gives .6439875, .426225, .6921375, and .6439875. Without T2SQL the nearest positive/negative modules have $F = .602075$ and .7822; their midpoint is .6921375. Replay uses equal module weights for value reporting and verifies that sign-calibrated thresholds maximize both native and normalized empirical values. Positive weighting does not alter optimal separation on these data. This confirms the archived T2SQL Retain/Discard/Discard route, not when candidates were frozen.

RQ3 averages normalized gains within each module, then equally across seven modules including defaulted SRR. Full/Discard/LOFO/full-data means are 22.11/21.90/31.09/33.71 percentage points; LOFO retains 92.2% of the full-data mean. These module averages evaluate routing and synthesis jointly, not the threshold alone.

Fixed-skill paired bootstrap. The bootstrap resamples paired evaluation results across methods, holding the constructed skills, diagnostic coordinates, and calibrated thresholds fixed. Each resample compares LOFO with the larger of the resampled Always Full and Always Discard values, using the RQ3 aggregation above: $\widehat{V}^{(b)}_{\mathrm{LOFO}} > \max\{\widehat{V}^{(b)}_{\mathrm{Full}}, \widehat{V}^{(b)}_{\mathrm{Discard}}\}$. This inequality holds in 97.8% of samples. The comparison reselects the stronger fixed baseline in each resample; it does not recalibrate thresholds, reconstruct skills, or substitute component scores for jointly executed CE/T2SQL skills. It measures evaluation uncertainty conditional on the fixed skills, not uncertainty from policy learning or skill construction. The aggregate replay verifies the module means separately from this bootstrap analysis.

Cost and baseline provenance. FE Compiler costs are \$3.09 on Taobao and \$2.47 on Dia (Table 16), giving the domain average $(3.09 + 2.47)/2 = 2.78$ in Table 1. Figure 8 projects $C(N) = C_0 + Nc$ for $N = 1$–1000, without an additional cache adjustment. In CE, CD, FE, and T2SQL order, the retained conversion charges are approximately \$7.905, \$2.358, \$5.500, and \$3.966 for Compiler, Full, and Discard. EvoSkill's iterative-construction charges are \$80.710, \$801.123, \$2083.517, and \$143.485. AdaSkill adds a fixed \$6.12 diagnostic allowance to each conversion charge; Source and Base incur no new construction charge. This allowance uses the CE 6×6 execution budget as a common accounting assumption, not a measurement of each family's complete diagnostic overhead. Source, Base, Compiler, EvoSkill, and AdaSkill execution rates follow Table 1; Full and Discard retain their matched-ablation rates, respectively $(0.223, 3.108, 6.177, 0.512)$ and $(0.211, 2.577, 4.846, 0.247)$ USD. Source FE rates remain operating-range estimates. The curves are projections under these fixed charges, rather than measurements at every deployment volume. For $c_B > c_A$, the real-valued crossover is $(C_{0,A} - C_{0,B})/(c_B - c_A)$; other signs require separate handling.

Scope of statistical evidence. Correlations are descriptive associations; conventional row-level p values do not account for all shared-workflow dependence or estimated diagnostic coordinates. Clustered regression intervals remain small-sample approximations with about eleven dataset clusters, not distribution-free guarantees. Complete run manifests are needed to reconcile binary denominators, diagnostic sampling counts, and final-result versions. Summary replay verifies internal arithmetic, not independent replication of agent execution.

B Experimental Design and Evaluation Details

B.1 Task and System Coverage

Text-to-SQL constructs and repairs programs over schemas; Causal Estimation (CE) selects and configures estimators; Causal Discovery (CD) searches graphs under causal constraints; and Feature Engineering (FE) generates and evaluates executable artifacts. These regimes vary output type, scorer, search structure, and source control while exposing the same identifiable objects: executable behavior, objective outcomes, and task-qualified pipeline intervention. The corresponding source workflows are APEX-SQL (Cao et al., 2026), CAIS (Verma et al., 2025), MATMCD (Shen et al., 2025), and FELA (Ouyang et al., 2025).

Each source is a published workflow selected for its reported target-task performance, supplying a substantive structural proposal rather than an arbitrary instruction set. Qualification maps APEX-SQL control to schema and executable-query metrics, CAIS to method selection and estimation error, MATMCD to graph F1, and FELA to downstream AUC; callable resources must also pass the target interface. These gates are fixed from source artifacts before any target skill is evaluated. In the matched intervention, Full and Discard retain the same validated tools and declarative knowledge and remove the same runtime orchestration; they differ only in whether task-qualified source-pipeline guidance is retained.

EvoSkill diagnoses training failures and iteratively rewrites a skill without decomposing source control or using F. EvoSkill is evaluated across CE, CD, T2SQL, and FE. FE compares Base Agent, EvoSkill, Workflow Compiler, and AdaSkill; FELA is the native source-workflow reference. Source Workflow rows retain their native orchestration and operating configurations. Appendix F follows the complete CAIS conversion from eight shared-state micro-tools to one skill.

Table 2: **Published, high-performing source workflows and externalized control.** Each source exposes an analytical decision loop through staged execution, specialized modules, shared artifacts, or feedback.

Task	Source	Control architecture	Externalized analytical decisions
T2SQL	APEX-SQL	Staged and parallel workflow modules	Hypothesis planning; dual-path schema pruning; parallel profiling; global synthesis; SQL exploration, execution, and verification.
CE	CAIS	Four-stage, eight-micro-tool pipeline	Query parsing; data analysis; variable interpretation; method selection; estimation; validation; refinement; reporting.
CD	MATMCD	Two modules in a centralized workflow	Retrieved-context augmentation; causal-constraint construction; statistical graph estimation and refinement.
FE	FELA	Role-specialized evolutionary workflow	Idea and code generation; criticism; evaluation; evolutionary selection over shared memory.

B.2 Metric Definitions

For CE, MSA measures agreement with the benchmark's canonical method label. MRE measures clipped relative deviation from the benchmark reference: $e_i = \min\{|\hat{\tau}_i - \tau_i|/|\tau_i|, 1\}$ for $\tau_i \neq 0$. The archived recomputation averages errors over available estimates with nonzero, nonmissing references. The diagnostic adapter scores a missing estimate as zero utility; for a zero reference, error is zero for a zero estimate and one otherwise. Paired comparison counts are reported in Table 10.

For T2SQL, EX scores executed results; schema metrics score the auxiliary `selected_columns` declaration. For an available declaration P_i and nonempty gold set G_i, define $r_i = |P_i \cap G_i|/|G_i|$ and $p_i = |P_i \cap G_i|/|P_i|$, with $p_i = 0$ for empty P_i. The per-question scores are $\mathrm{SRR}_i = \mathbf{1}[P_i \supseteq G_i]$ and $f_i = 2p_i r_i/(p_i + r_i)$, with $f_i = 0$ when $p_i + r_i = 0$. The archived evaluator averages r_i, p_i, f_i over this eligible set to obtain NSR, NSP, and NSF; missing declarations and empty gold sets are excluded. EX includes all records, counting execution failures as incorrect. CD uses edge F1; FE uses test AUC.

B.3 Data Separation and Inference

Diagnosis, conversion, and final evaluation use disjoint data roles fixed before skill synthesis. F is computed from repeated Base Agent runs on diagnostic questions, using gold metric scores only to construct outcome-distance ranks. CE uses six diagnostic questions per dataset; T2SQL uses ten; both execute six runs per question. Final CE results are scored on locked Textbook, Synthetic, and Real sets with 39, 45, and 31 valid cases. Final T2SQL results use 147 BIRD and 120 Spider questions; the suffixes in BIRD-147 and Spider-120 denote these evaluation-set cardinalities. No final-evaluation instance or answer enters diagnosis or conversion.

For per-question tasks, F aggregates identifiable diagnostic questions whose observed scores vary across runs. The diversity planner expands coverage across task-valid strategies, so the probe measures exploratory behavioral support. It returns only metric-level F values to the converter; probe traces, question identities, and answers are never conversion inputs. The converter receives the source paper, workflow code, task specification, and tool interfaces, but no benchmark instances or answers. Threshold calibration minimizes sign errors on aggregated development modules and, on these data, also maximizes Eq. (45) (Appendix A.10). The resulting rule is frozen before evaluating the excluded target regime. The protocol prescribes freezing before target execution; the aggregate replay does not independently verify access histories.

Diagnosis uses all within-question pairs: six runs yield $\binom{6}{2} = 15$, or 90 across six CE questions before validity and identifiability filtering. CD and FE performance tables report independent execution means and standard deviations; CE and T2SQL preserve question identities for paired uncertainty. The atomic and matched intervention analyses use dataset–metric outcomes as their inferential units and never count run pairs as independent samples. Workflow-regime fixed effects absorb level differences, and dataset-clustered inference respects metrics that share executions. Sonnet 4.6 supplies the controlled intervention study. A complete replication of the task and dataset suite, Base

conditions, and skill conditions on GPT-5.1 tests whether the system-level ordering survives a new backbone (Appendix C.3).

C Behavior-Outcome Freedom

C.1 Computation Details

For each (task, dataset, metric) tuple, we compute freedom from two complementary perspectives: F_{out} (structured output distances) and F_{trace} (embedding-based process-trace distances). Both are computed from the "Base Agent" condition before any skill is generated or evaluated. The agent, harness, planner, and sampling procedure jointly define the measured support.

Table 3: Evaluator-facing state H and diagnostic projection $B = \phi(H)$ used for F_{out}. The fixed harness scores H; a coarse B is not assumed outcome-sufficient and is covered by $R_{\pi,B}$. Trace-space F uses response-free embeddings and scaled cosine dissimilarity.

Task	**Scored state H**	**Diagnostic B**	$d(B_i, B_j)$	**Aggregation**
CD	Submitted directed graph	Edge set E	$1 - \text{Jac}(E_i, E_j)$	Dataset
CE	Method, configuration, estimate	Method; execution audit	$\mathbf{1}[m_i \neq m_j]$; hierarchical	Question mean
FE	Executable transform and evaluated features	Feature set $\mathcal{F}$	$1 - \text{Jac}(\mathcal{F}_i, \mathcal{F}_j)$	Dataset
T2SQL	SQL and schema declaration	Metric-scoped token/set view	$1 - \text{Jac}(B_i, B_j)$	Question mean

Metric–representation overlap. Method-class distance and MSA share a label definition. For binary pair variables $U = \mathbf{1}[m \neq m']$ and $V = |\text{MSA} - \text{MSA}'|, V \leq U$. If $a = \Pr(U = 1)$ and $b = \Pr(V = 1)$ are nondegenerate, $\rho_S(U, V) = \sqrt{b(1-a)/(a(1-b))} \geq 0$. Low canonical MSA Freedom is therefore partly mechanical, not independent evidence of beneficial pipeline transfer. Response-free traces without method labels and matched interventions provide separate checks.

Representation contract. The scored state H contains the artifact and configuration consumed by the fixed evaluator; no gold outcome is part of the agent-visible behavior. CD and T2SQL scores are deterministic given the submitted graph or query and fixed evaluation instance, while FE uses a fixed downstream harness. CE additionally depends on the numerical estimate and configuration, so estimator identity is explicitly only a diagnostic projection. Consequently, the theory conditions the evaluator kernel on H and assigns any intervention-dependent change left inside a canonical B cell to $R_{\pi,B}$. The reported canonical and trace statistics test whether the ordering is stable across projections; they are not used to assert that every coarse projection has zero remainder.

Planner design versus the iid theory. For a fixed list of N hints inducing run laws $P_1, \ldots, P_N$, non-diagonal pairs follow

$$Q_{\text{design}} = \frac{1}{N(N-1)} \sum_{i \neq j} P_i \otimes P_j, \qquad \bar{P} \otimes \bar{P} = \frac{1}{N^2} \sum_{i,j} P_i \otimes P_j.$$

The laws differ: the second includes two independent executions of the same hint. Equation (2) is computed on the finite design; its ranks are design-specific. The iid mechanism theorems do not automatically certify this diagnostic. No-planner controls test ordering robustness but do not prove equality of these laws across task families. Run resampling conditional on a hint list does not capture planner uncertainty; a full design-based audit needs repeated planner lists and their run manifests.

Diversity planner. Low-temperature inference often concentrates on a small set of dominant outputs, so each question is seeded by a lightweight diversity planner. The planner is the diagnostic intervention used to expose task-valid alternatives whose suppression a structural skill may cause. It is one LLM call using the same backbone and receives the original problem to generate N methodologically distinct solution strategies, such as different statistical methods or schema traversal orders. Each strategy is prepended as a brief hint, and the base agent executes independently for every hint. Conditional execution independence does not make the distinct-hint runs identically distributed. The reported diagnostic is conditional on this planner design, not an iid estimate of Eq. (1). The prompt requires methodologically sound approaches and permits fine-grained variations, including different covariate sets within one method. No additional sampling temperature or top-p manipulation is used. The full prompt template follows.

Freedom probe: diversity-planner prompt

```
You are given a problem below. Generate {N} reasonable approaches to solve it.

Requirements:
1. Every approach MUST be methodologically sound and appropriate for the
   problem -- do not invent approaches just to fill the list
2. Approaches may be similar to each other; fine-grained differences are
   acceptable (e.g., same method with different covariate sets, different
   hyperparameters, or slightly different model specifications)
3. If the problem only supports a small number of truly reasonable
   strategies, generate variations within those strategies; do not
   force unrelated ones
4. Be specific about WHAT to do: name the method, key steps, and any
   important implementation choices. Focus on the high-level methodology
   and strategy, not specific values, feature names, or implementation
   details (those are for the executor to decide)
5. Each approach should be self-contained and actionable

Output format (use exactly this format):
## Approach 1: [Brief Name]
[2-4 sentences describing the core idea and key steps]

## Approach 2: [Brief Name]
...

## Original Problem

{problem}

**IMPORTANT: YOUR TASK RIGHT NOW:**
Do NOT solve the problem above. Do NOT output JSON, SQL, a matrix, or
any final answer. Your ONLY job is to generate {N} reasonable approaches
using the format at the top of this prompt.
```

Each generated approach is then prepended to the original problem as a one-line hint (`Hint: {approach}`) before the base agent executes. This planner is the only LLM prompt introduced by the Freedom probe. Output and trace extraction, distance construction, the Mantel statistic, aggregation, and bootstrap resampling are deterministic computations.

F_{out} computation. We instantiate Eq. (2) with $\mathcal{X}$ = output space. The behavioral distance $d(\cdot, \cdot)$ is task-specific (Table 3).

F_{trace} computation. We instantiate Eq. (2) with $\mathcal{X}$ = trace embedding space. Each process trace τ_i is extracted from verbose run output and contains intermediate assistant messages, tool calls, and tool results. The final response is removed before the trace is embedded with `text-embedding-3-large` using persistent disk caching. The behavioral dissimilarity is $d(\tau_i, \tau_j) = (1 - \cos_{\text{sim}}(\text{embed}(\tau_i), \text{embed}(\tau_j)))/2$. This positive scaling does not change ranks or F. Thus F_{trace} measures process variation, while F_{out} separately measures structured final outputs. Runs where embedding fails are dropped from the pairwise computation.

Bootstrap uncertainty summaries. For dataset-level estimates, we resample runs and reconstruct both pairwise matrices before recomputing F (B=1000); this preserves the dependence among pairwise entries that share a run. For per-question metrics (CE, T2SQL), we first compute F per identifiable question and then bootstrap questions while recomputing the canonical mean used by both routing and cross-task analysis. Some cells contain few identifiable questions, so the resulting standard errors are descriptive and are not presented as calibrated confidence intervals.

Exclusion criteria and default routing. For per-question tasks (CE, T2SQL), we include only *mixed questions* where scores vary across the observed runs, because otherwise the sample correlation denominator is zero. The diagnostic therefore has two reported outputs: the identifiable fraction K/M and the conditional coupling estimate over those K questions. Observed uniformity does not prove zero population variance; rare-varying questions may be omitted when n is small. For per-dataset tasks (CD, FE), dissimilarities are computed globally over the available runs. A per-question metric must yield at least three identifiable questions to justify pipeline retention. CE-Synthetic-MSA and the sparse Spider NSF/SRR estimates remain audit entries. The evaluated T2SQL skill

was routed by BIRD EX, which has four identifiable questions. For F_{trace}, runs where trace embedding fails are dropped from the pairwise computation. Routing uses F_{out} whenever it reaches the minimum and falls back to F_{trace} only when the output statistic is undefined. If neither representation reaches the minimum after additional sampling, the system takes the default Discard route.

This fallback retains the validated capability layer and emits no mandatory source-pipeline clauses. Table 4 reports the complete finite-sample audit.

Table 4: Finite-sample identifiability audit for per-question F. M is the number of sampled questions; K_{var} has observed score variation; K_{out} and K_{trace} yield identifiable statistics in the two representations. Mean score variance is computed across K_{out} on a $[0, 1]$ score scale. Values with $K < 3$ are audit entries and cannot independently determine a route.

Task	**Dataset**	**Metric**	M	K_{var}	K_{out}	K_{trace}	**Mean score var.**
CE	Textbook	MSA	6	6	6	6	0.233
CE	Textbook	MRE	6	6	6	6	0.114
CE	Real	MSA	6	5	5	5	0.300
CE	Real	MRE	6	6	5	6	0.136
CE	Synthetic	MSA	6	1	1	1	0.300
CE	Synthetic	MRE	6	6	6	6	0.039
T2SQL	BIRD-147	EX	10	4	4	4	0.275
T2SQL	BIRD-147	NSF	10	6	6	6	0.006
T2SQL	BIRD-147	SRR	10	2	2	2	0.250
T2SQL	Spider-120	EX	10	2	2	2	0.250
T2SQL	Spider-120	NSF	10	1	1	1	0.029
T2SQL	Spider-120	SRR	10	0	0	0	–

All score-varying questions produced valid trace embeddings, so the trace embedding failure count is zero. The sole representation mismatch is CE Real MRE: one question has varying scores but a constant output representation, leaving $K_{\text{out}} = 5$ and $K_{\text{trace}} = 6$. Each T2SQL dataset has 60 diagnostic executions (10×6). The underlying runs contain no execution failures for CE or BIRD and three failures among 60 Spider runs; the statistics use the remaining valid runs.

C.2 Operational AdaSkill Protocol

Mechanical component boundary. Classification is clause-level and functional. *Knowledge* is a declarative proposition about the domain, data schema, estimator assumptions, or tool semantics. It cannot prescribe, prefer, prohibit, order, repeat, or terminate an action. *Pipeline* is any mapping from an observed state to an action policy, including soft recommendations, hard requirements, prohibitions, numbered procedures, conditional branches that select tools or methods, validation gates, retries, and stopping rules. If a clause mixes a fact with an action rule, the entire clause is pipeline. *Orchestration* additionally requires runtime control across modules or agents through handoffs, shared state, voting, debate, or loops. This rule is independent of section names and surface wording: placing "FORBIDDEN" or "do this first" under a knowledge heading does not make it knowledge.

AdaSkill converter contract. Every reported conversion follows the same five-step contract. First, the converter inventories callable tools, declarative knowledge, state-to-action pipeline clauses, and orchestration mechanisms from the source paper and code. Second, it retains tool interfaces, rewrites knowledge as descriptive reference material, and removes agent identities, handoffs, shared-state operations, and runtime control loops. Third, it applies the empirical-value threshold from Section 3.4: retain all task-qualified pipeline clauses for $F < t$ and remove them for $F \geq t$. In a family-excluded fold, Eq. (45) is optimized once on the other regimes, and the threshold is then immutable for every metric in the target family. Fourth, it emits one Markdown skill with fixed sections for scope, tools, reference knowledge, metric-specific guidance, and failure checks. Fifth, an automated clause audit checks imports, section presence, binary routing compliance, and the absence of benchmark instances or answers. Headings do not override the functional classification, and a mixed fact–directive clause inherits the pipeline label. The converter receives metric-level F values and the already frozen threshold but never receives probe traces, question identities, skill outcomes, or final-test feedback.

AdaSkill Converter: binary workflow-to-skill prompt

```
You are the AdaSkill workflow-to-skill converter. Convert a task-qualified source
workflow into one executable skill for a tool-using agent.

Inputs:
  source paper: {paper_path}
  source code: {codebase_path}
  task specification: {task_specification}
  tool interfaces: {tool_interfaces}
  metric-level Freedom values: {metric_freedom}
  frozen threshold: {threshold_t}
  output directory: {output_dir}

1. Inventory every source clause by function:
     tool = callable implementation;
     knowledge = declarative domain, data, estimator, or tool fact
                 with no action directive;
     pipeline = any recommendation, prohibition, ordered step, state-to-action
branch, validation gate, retry, or stopping rule;
     coordination = runtime handoff, shared state, voting, debate, or control loop.
   Classify every mixed fact-and-directive clause as pipeline. Section headings and
polite wording do not change the class.
2. Preserve callable tools. Rewrite relevant knowledge as descriptive reference
material. Remove agent identities, handoffs, shared state, voting, debate, and
runtime control loops.
3. For each metric module, apply one binary structural rule:
   use the supplied frozen threshold t = {threshold_t};
   if F < t: retain the task-qualified pipeline clauses;
   if F >= t: remove all pipeline clauses.
   Do not infer a new threshold and do not create an intermediate route.
4. Compose the routed metric modules into one skill. Deduplicate shared tools and
knowledge. Keep conflicting metric requirements explicitly scoped to their
metric-specific decisions.
5. Write SKILL.md with fixed sections for scope, tools, reference knowledge,
metric-specific guidance, and failure checks. Write required tool files under tools/.
6. Reclassify every emitted clause by the same functional rule. On a discard route,
reject any source-derived state-to-action clause even if it appears under reference
knowledge. Validate tool imports, required sections, binary routing compliance, and
the absence of benchmark instances or answers.

Do not inspect probe traces, question identities, skill outcomes, validation
feedback, or final-test data. Write the complete skill package directly to
{output_dir}.
```

Multi-metric composition. Metric modules are composed by the decision they score, not by averaging their F values. In CE, MSA controls method-selection guidance and MRE controls numerical-estimation guidance. In T2SQL, SRR and NSF govern schema-evidence checks, while EX governs executable-query construction and repair. Shared tools and knowledge are deduplicated. A retained pipeline remains scoped to its metric-specific decision and cannot impose structure on a subproblem assigned to the discard route. This fixed mapping produces one executable skill from multiple metric-conditioned modules.

Structural robustness and dependence control. The atomic sweep contains 15 Pipeline-only interventions against the Base Agent and excludes the Dia Workflow Compiler proxy. Its normalized association is Pearson $r = -0.623$ ($p = 0.013$) and Spearman $\rho = -0.568$ ($p = 0.027$). This estimate uses neither the Full nor Discard outcomes.

The complete matched estimand uses all 16 capability-matched interventions. Because every metric has already been mapped to higher-is-better utility in $[0, 1]$, its primary effect is the direct utility difference $u_{\text{full}} - u_{\text{discard}}$. Workflow-regime fixed effects estimate a slope of -9.18 utility points per unit F. Clustering shared-metric dependence at the dataset level gives a 95% interval of $[-13.41, -4.94]$ and $p < 0.001$. Native utility differences give Pearson $r = -0.798$ ($p = 0.0002$), and the normalized effect gives $r = -0.673$ ($p = 0.0043$). The strict route-grade subset requires $K \geq 3$ for per-question diagnostics and retains 13 interventions; it recovers slope -9.72, interval $[-14.00, -5.45]$, and $p < 0.001$. Leave-one-regime fixed-effect slopes remain negative after removing CE, CD, T2SQL, or FE. Baseline utility is unrelated to normalized treatment effect ($p = 0.90$).

For CE, we additionally refine the task-authored representation without an embedding model. The hierarchical execution signature keeps estimator identity as the coarse decision and separates executions within an estimator by statistical packages, estimator calls, dataframe columns, causal-design tokens, keyword arguments, and whitelisted configurations such as bandwidth, cutoff, kernel, and estimand. It parses assistant tool calls only. Tool results, final responses, and estimated effects are never read. Table 5 shows the Textbook scorer reversal under both representations: MSA/MRE are $0.098/1.012$ for method identity and $0.181/1.085$ for execution signatures. Real gives $0.236/0.697$ and $0.255/0.762$, respectively. For Real MRE, $K_{\text{method}} = 5$ and $K_{\text{exec}} = 6$: the comparison includes a change in question eligibility, not representation alone. Replacing the five canonical CE coordinates by the reported three-decimal structured values preserves the matched ordering on the route-grade set (normalized $r = -0.676$, $p = 0.011$; utility-point $r = -0.788$, $p = 0.0014$) and complete audit (normalized $r = -0.679$, $p = 0.0038$; utility-point $r = -0.793$, $p = 0.0002$). This sensitivity analysis uses each representation's eligible questions; it is not a common-question comparison or an estimate of $R_{\pi,B}$.

Table 5: **CE execution-signature audit.** The canonical selected-estimator distance is enriched hierarchically with response-free execution configuration parsed from retained tool calls. Different estimator classes have distance one; within-class distance is one half the Jaccard distance over statistical packages, estimator calls, dataframe columns, causal-design tokens, keyword arguments, and whitelisted numeric configurations. Tool results, final responses, and estimated effects are excluded. K_{method} and K_{exec} count identifiable questions for each representation.

Dataset	**Metric**	K_{method}	K_{exec}	F_{method}	F_{exec}
Textbook	MSA	6	6	0.098	0.181
Textbook	MRE	6	6	1.012	1.085
Real	MSA	5	5	0.236	0.255
Real	MRE	5	6	0.697	0.762
Synthetic	MRE	6	6	0.638	0.651

Across the 16 matched contrasts, the normalized effect gives Pearson $r = -0.673$ ($p = 0.0043$) and Spearman $\rho = -0.735$ ($p = 0.0012$); native-score differences give $r = -0.798$ ($p = 0.0002$) and $\rho = -0.747$ ($p = 0.0009$). Seven metric means give native-scale $r = -0.855$, and every regime deletion remains negative. Table 7 reports the complete treatment audit.

Threshold-free ranking. Treating Full $>$ Discard as the positive structural action, $-F$ attains ROC AUC 1.000 across all 16 interventions and on the strict-identifiability subset. Every leave-one-regime-out deletion preserves perfect ordering. AUC is descriptive: row-wise label exchangeability is not justified when metrics share executions, so we do not interpret label enumeration as an exact significance test.

The ordering is also insensitive to effects near zero. Deleting one through six smallest Full-minus-Discard contrasts preserves ROC AUC 1.000 on the remaining points. This sensitivity check does not incorporate estimation error in F or treatment signs. Table 6 summarizes both ranking audits.

Table 6: **Threshold-free ranking of structural value.** The positive class is Full $>$ Discard. Scores use $-F$. AUC is a sample ranking, not an exact test under shared-execution dependence.

Audit	n	Full $>$ Discard	Discard $>$ Full	ROC AUC
Route-grade	13	5	8	**1.000**
Complete	16	5	11	**1.000**

Workflow-regime fixed effects describe within-regime effect modification, and dataset clustering keeps metrics sharing executions in the same dependence block. On all 16 interventions, the utility-point slope is -9.18 with 95% interval $[-13.41, -4.94]$ ($p < 0.001$). The strict-identifiability subset gives slope -9.72, interval $[-14.00, -5.45]$ ($p < 0.001$). Native, normalized, and every regime-deletion analysis preserve the negative direction.

Native-score effects, metric aggregation, family deletion, and the baseline-utility control all preserve the structural ordering. These tests target the matched intervention itself and remain distinct from AdaSkill's cross-fitted operating policy.

Table 7: **Complete matched pipeline interventions** Base, Full, and Discard are higher-is-better utilities; MRE is converted as $1 - \mathrm{MRE}$. The final column is the normalized Full-minus-Discard treatment effect. Grey rows have fewer than three identifiable questions.

Task	Dataset	Metric	K	F	Base	Full	Discard	τ_{pipe}
CE	Textbook	MSA	6	.098	.615	.846	.769	+20.0
CE	Real	MSA	5	.235	.871	.903	.871	+25.0
CE	Textbook	MRE	6	1.012	.688	.737	.769	-10.3
CE	Real	MRE	5	.697	.817	.827	.831	-2.1
CE	Synthetic	MRE	6	.638	.894	.924	.951	-26.0
CD	AutoMPG	F1	–	.914	.670	.652	.667	-4.5
CD	Child	F1	–	.291	.803	.952	.918	+17.3
CD	DWDClimate	F1	–	.605	.558	.701	.663	+8.6
CD	Sachs	F1	–	.598	.869	.924	.893	+23.7
T2SQL	BIRD-147	EX	4	.686	.592	.660	.673	-3.3
T2SQL	BIRD-147	NSF	6	.816	.905	.925	.934	-9.3
T2SQL	BIRD-147	SRR	2	.697	.718	.782	.791	-3.1
FE	Taobao	AUC	–	1.037	.660	.657	.670	-3.8
T2SQL	Spider-120	EX	2	1.071	.600	.642	.667	-6.2
T2SQL	Spider-120	NSF	1	.787	.114	.585	.597	-1.4
FE	Dia	AUC	–	1.337	.804	.798	.812	-7.0

Baseline test utility is the locked Base-Agent result reported in the matched intervention table and is unrelated to the structural effect ($r = 0.033$, $p = 0.904$). An unsigned distance-dependence statistic corroborates the geometric premise but discards whether the association is concordant or anti-concordant. The signed rank construction in F supplies the orientation of the behavior–outcome geometry; pipeline direction remains intervention-specific.

Table 8: **Dataset-level audit of cross-fitted routing and the full-information structural oracle.** Base, Full, and Discard are matched candidates. LOFO selects a candidate for single-metric CD/FE (Selected); multi-metric CE/T2SQL directly executes a jointly composed skill (Composed). The descriptive in-sample oracle chooses the better Full/Discard outcome after evaluation; it is not an AdaSkill execution. Values are native percentages; MRE is lower-is-better. Discard MSA on Textbook (30/39) and EX on Spider (80/120), and LOFO EX on BIRD (98/147) and Spider (79/120), are displayed to one decimal; calculations use unrounded proportions.

Family	Dataset	Metric	Base	Full	Discard	Route	Output type	LOFO policy	In-sample oracle
CE	Textbook	MSA	61.54	84.62	76.9	P	Composed	89.70	Full (84.62)
CE	Real	MSA	87.10	90.32	87.10	P	Composed	96.80	Full (90.32)
CE	Textbook	MRE	31.25	26.27	23.06	D	Composed	23.10	Discard (23.06)
CE	Real	MRE	18.33	17.33	16.95	D	Composed	15.90	Discard (16.95)
CE	Synthetic	MRE	10.56	7.63	4.88	D	Composed	4.30	Discard (4.88)
CD	AutoMPG	F1	67.00	65.20	66.67	D	Selected	66.67	Discard (66.67)
CD	Child	F1	80.30	95.20	91.80	D	Selected	91.80	Full (95.20)
CD	DWDClimate	F1	55.80	70.10	66.30	D	Selected	66.30	Full (70.10)
CD	Sachs	F1	86.90	92.40	89.30	D	Selected	89.30	Full (92.40)
T2SQL	BIRD-147	EX	59.20	65.99	67.35	P	Composed	66.7	Discard (67.35)
T2SQL	BIRD-147	NSF	90.50	92.52	93.40	D	Composed	93.80	Discard (93.40)
T2SQL	BIRD-147	SRR	71.80	78.23	79.10	D	Composed	80.20	Discard (79.10)
FE	Taobao	AUC	66.02	65.70	67.00	D	Selected	67.00	Discard (67.00)
T2SQL	Spider-120	EX	60.00	64.20	66.7	P	Composed	65.8	Discard (66.7)
T2SQL	Spider-120	NSF	11.40	58.50	59.70	D	Composed	60.20	Discard (59.70)
FE	Dia	AUC	80.40	79.82	81.20	D	Selected	81.20	Discard (81.20)

C.3 Backbone Replication on GPT-5.1

We repeat the complete dataset suite and both Base and skill-augmented conditions with GPT-5.1 under the same skill definitions, prompts, tools, and evaluation protocol. Figure 11 shows that the system-level relationship survives the changed behavior law. Output-space F gives $r = -0.71$ ($p < 0.01$), and response-free trace-space F gives $r = -0.79$ ($p < 0.001$).

C.4 Question-Level Uncertainty for Final CE and T2SQL Results

Table 9 reports uncertainty over evaluation questions. Wilson intervals are used for MSA and EX. MRE intervals use 10,000 question-level bootstrap samples. Table 10 reports paired question-bootstrap intervals for direct method gains where retained outputs permit question matching. CE Overall MRE averages dataset-level paired mean differences with equal weights; its stratified boot-

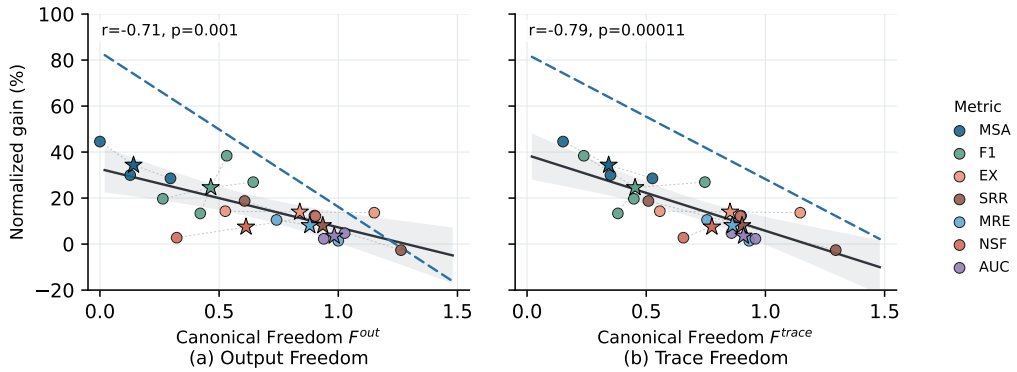


Figure 11: **The pre-skill coupling signal persists under GPT-5.1.** Canonical output-space **(a)** and response-free trace-space **(b)** estimates preserve the negative ordering with AdaSkill gain ($r = -0.71$ and $r = -0.79$) across all 18 dataset–metric tuples. Both Freedom and gain are re-estimated under GPT-5.1 using the same evaluation protocol as Figure 3. Black solid lines fit GPT-5.1; blue dashed lines transfer the Sonnet 4.6 fits from Figure 3 without refitting. Circles are individual datasets; stars are metric means.

strap resamples questions independently within each dataset before averaging. Synthetic Acc@τ thresholds each question at clipped relative error $\leq \tau$ and bootstraps the paired binary differences.

The final BIRD EX display retains the two-stage rounding of its percentage, $67.3469\ldots \rightarrow 67.35 \rightarrow 67.4$; its Wilson interval uses the underlying binary count rather than the rounded percentage. LOFO EX uses the unrounded proportions specified in Table 8.

Table 9: Final AdaSkill estimates with 95% question-level intervals. MRE is lower-is-better; all entries are percentages.

Task	**Dataset**	**Metric**	**Estimate**	**95% interval**	n
CE	Textbook	MSA	89.7	$[76.4, 95.9]$	39
CE	Textbook	MRE	23.1	$[14.7, 32.4]$	39
CE	Synthetic	MSA	100.0	$[92.1, 100.0]$	45
CE	Synthetic	MRE	4.3	$[2.7, 6.1]$	45
CE	Real	MSA	96.8	$[83.8, 99.4]$	31
CE	Real	MRE	15.9	$[8.2, 25.4]$	31
T2SQL	BIRD-147	EX	67.4	$[59.4, 74.4]$	147
T2SQL	Spider-120	EX	66.7	$[57.8, 74.5]$	120

C.5 Sensitivity to M and N

A central design decision in computing F is how many questions (M) and how many runs per question (N) to collect. Both quantities directly affect the reliability of the Mantel-test estimate: too few runs leave pairwise distances noisy; too few questions make the question-level aggregate unstable. At the same time, collecting $M \times N$ runs has a real cost. Figure 9 uses the measured average of \$0.17 per run, so cost is linear in $M \times N$ and reaches \$20.4 at the maximum budget of M=10, N=12.

Figure 9 reports F_{out}, while Figure 12 reports the complementary F_{trace} sweep and planner ablation. These are different behavioral representations and need not coincide numerically at the same (M, N) operating point. The trace-space sweep reveals a clear saturation pattern for both $F^{\text{trace}}_{\text{MSA}}$ and $F^{\text{trace}}_{\text{MRE}}$. For $F^{\text{trace}}_{\text{MSA}}$ (panel a), the estimate stabilises once $N \geq 5$: beyond this point the colour map changes by less than two percentage points per additional run column. For $F^{\text{trace}}_{\text{MRE}}$ (panel b), even N=3 places the estimate above the calibrated threshold range; additional runs reduce variation without changing the binary decision. In both panels, increasing M beyond 6 produces only marginal shifts across the grid, identifying M=6 as the operating point.

Taken together, the gold operating point at M=6, N=6 (\$6.12) is the empirical knee of the accuracy–cost curve, consuming 30% of the maximum budget while preserving the binary separation. Relative

Table 10: **AdaSkill gains, aggregate robustness, and efficiency.** Positive values favor AdaSkill. Brackets give 95% intervals; bold intervals exclude zero. MRE gains are baseline minus AdaSkill, with relative reductions in parentheses. All CE comparisons and T2SQL comparisons with Base and Compiler use 200,000 paired question-bootstrap samples. CE Overall MRE uses a dataset-stratified paired bootstrap with equal dataset weights. Synthetic MSA intervals degenerate because all paired outcomes are successes; they do not establish population equivalence. Ranges in n denote pair-specific complete-case counts across baseline comparisons. Acc@τ is the fraction of Synthetic questions with relative error at most τ. Efficiency entries are signed savings relative to each baseline: positive values are reductions and negative values are increases. T2SQL EvoSkill entries marked † use Newcombe intervals from final counts.

Task	Dataset and metric	n	vs. Base	vs. Compiler	vs. EvoSkill
Panel A: Dataset-level performance gains (percentage points)					
CE	Textbook MSA	39	**28.2 [15.4, 43.6]**	7.7 [−2.6, 18.0]	**17.9 [2.5, 33.3]**
CE	Textbook MRE	38–39	8.1 [−0.8, 18.0] (26.0%)	3.2 [−4.4, 11.7] (12.2%)	2.3 [−4.6, 10.1] (9.1%)
CE	Synthetic MSA	45	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
CE	Synthetic MRE	45	**6.3 [1.1, 13.2]** (59.4%)	**8.6 [1.9, 17.0]** (66.7%)	2.8 [−1.0, 7.8] (39.4%)
CE	Real MSA	31	9.7 [−3.2, 22.6]	6.5 [−6.4, 19.4]	3.2 [−6.5, 12.9]
CE	Real MRE	31	2.4 [−6.3, 11.4] (13.1%)	1.4 [−12.1, 15.9] (8.1%)	5.7 [−5.2, 17.6] (26.4%)
T2SQL	BIRD-147 EX	147	**8.2 [3.4, 13.0]**	0.0 [−4.1, 3.4]	1.4 [−9.3, 12.0]†
T2SQL	Spider-120 EX	120	6.7 [−1.6, 15.0]	**9.2 [2.5, 15.9]**	4.2 [−7.9, 16.0]†
Panel B: CE aggregate and threshold robustness					
CE	Overall MRE	3 datasets	**5.6 [1.0, 10.5]** (28.0%)	4.4 [−1.4, 10.5] (23.4%)	3.6 [−1.0, 8.5] (20.0%)
CE	Synthetic Acc@10%	45	**13.3 [4.4, 24.4]**	**11.1 [2.2, 20.0]**	2.2 [−6.7, 11.1]
CE	Synthetic Acc@20%	45	**11.1 [2.2, 20.0]**	**11.1 [2.2, 20.0]**	6.7 [−2.2, 15.6]
CE	Synthetic Acc@50%	45	4.4 [0.0, 11.1]	**8.9 [2.2, 17.8]**	4.4 [0.0, 11.1]
Panel C: CE efficiency and measured Pareto dominance					
CE	Performance cells W/T/L	6	5/1/0	5/1/0	5/1/0
CE	Inference-time saving	–	−52.9%	54.8%	21.2%
CE	Inference-cost saving	–	30.1%	66.0%	23.3%
CE	Pareto dominant in performance and cost	–	✓	✓	✓

to $M{=}N{=}10$ ($17.00), it reduces measurement cost by $2.8\times$. Cells with $N < 3$ are greyed out because fewer than three runs yield too few pairwise entries for the rank statistic to be informative.

Diversity planner ablation. Figure 12 compares trace-space F estimation with and without the diversity planner. Without it, each of the N runs is an independent random execution of the base agent. Because low-temperature inference concentrates on a small number of dominant reasoning paths, many runs produce near-identical trajectories and undersample plausible alternatives. At $N{=}6$, the no-planner $F^{\text{trace}}_{\text{MSA}}$ estimate is 42 versus the large-budget reference of approximately 53, whereas the planner-conditioned estimate is 52. At $M{=}N{=}20$, the two protocols recover nearly identical metric ordering and values. The planner therefore accelerates coverage of exploratory support without manufacturing the scorer reversal.

C.6 Freedom across Diagnostic Questions

Figure 13 shows every retained identifiable question-level estimate.

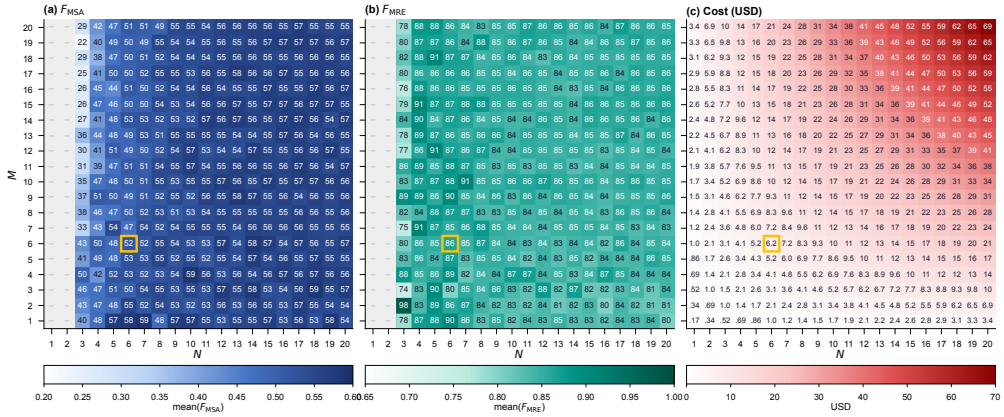
Each dot is an identifiable question-level F^{out} estimate from the retained diagnostic run for a dataset. The canonical dataset–metric coordinate is their mean, used without alteration in both structural analysis and routing. Budget-level stability is evaluated separately in Figures 9 and 12, while Table 4 reports how many questions identify each metric.

C.7 F Robustness across Datasets

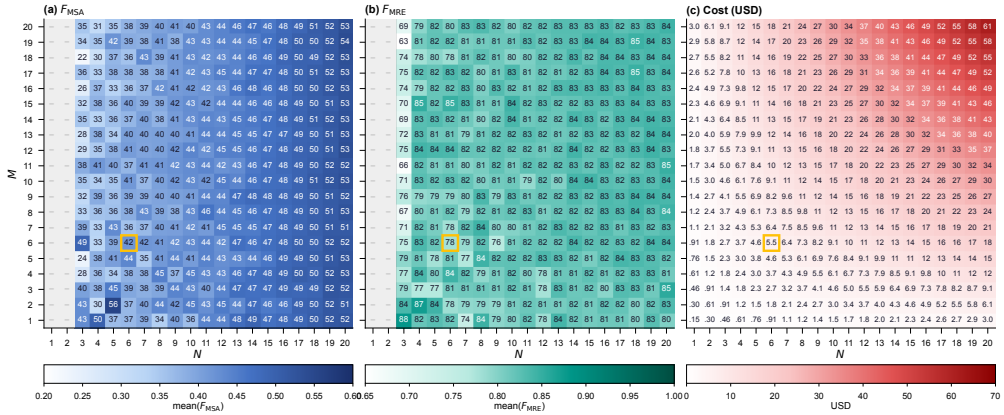
We test cross-dataset stability using Causal Estimation (CE), three datasets (Textbook, Real, Synthetic), and two metrics (MSA, MRE). The retained Synthetic-MSA diagnostic has only one identifiable question and is excluded from this distributional comparison, matching the identifiability audit in Table 4.

Tables 11 and 12 report, respectively, the per-dataset F distributions and the pairwise cross-dataset comparisons for both F^{out} and F^{trace}.

Distribution. Table 11 shows stable ordering alongside numerical variation. The identifiable Textbook and Real MSA estimates are close in both representations. For MRE, F^{out} ranges from 0.638



(a) F_{trace} with diversity planner (used in all experiments).



(b) F_{trace} without diversity planner (independent random runs).

Figure 12: **Diversity planner ablation for trace-space Freedom.** Both panels show $F^{\text{trace}}_{\text{MSA}}$ and $F^{\text{trace}}_{\text{MRE}}$ ($\times 100$) as a function of evaluation budget (M questions, N runs). At $M{=}N{=}20$, planner and no-planner protocols give similar displayed estimates ($F^{\text{trace}}_{\text{MSA}}$: 55 vs. 53; $F^{\text{trace}}_{\text{MRE}}$: 85 vs. 85). At $M{=}N{=}6$, the planner-conditioned values are closer to their respective large-budget values than the independent-run values. Figure 9 reports a separate output-space sweep.

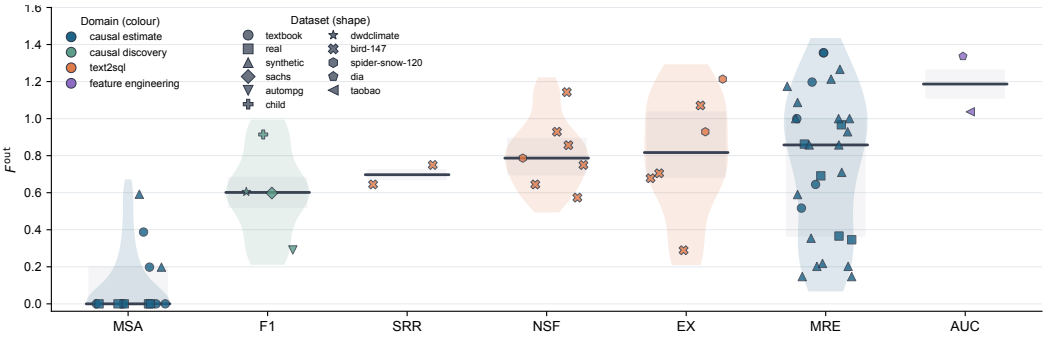


Figure 13: Question-level F^{out} values from the retained diagnostic run for each dataset. Colour encodes domain and shape encodes dataset.

to 1.012, while Real and Synthetic are closer (0.697 vs. 0.638, $\Delta\mu = 0.059$). These results preserve the MSA–MRE separation across the identifiable CE comparisons.

Cross-dataset comparison. Table 12 reports pairwise $\Delta\mu$ (absolute difference of means), 95% interval overlap, and exploratory Mann–Whitney U test p-values (no shared question IDs exist across datasets). The two valid MSA representation comparisons have interval overlap. For MRE/F^{out}, Real and Synthetic have $\Delta\mu = 0.059$ and $p = 1.00$. MRE/F^{trace} for Textbook versus Synthetic gives $p = 0.026$ with non-overlapping intervals, localizing the main cross-dataset difference to this representation and metric.

Table 11: F distribution per dataset for Causal Estimation. "n" = number of questions with a valid F estimate; std and var computed with Bessel's correction.

Dataset	Metric	Type	n	μ	σ	σ^2	Range
Textbook	MSA	F^{out}	6	0.098	0.163	0.027	[0.000, 0.388]
		F^{trace}	6	0.538	0.370	0.137	[0.147, 1.000]
	MRE	F^{out}	6	1.012	0.361	0.130	[0.517, 1.355]
		F^{trace}	6	1.109	0.268	0.072	[0.780, 1.491]
Real	MSA	F^{out}	5	0.236	0.256	0.066	[0.000, 0.592]
		F^{trace}	5	0.559	0.391	0.153	[0.147, 1.000]
	MRE	F^{out}	5	0.697	0.367	0.135	[0.147, 1.167]
		F^{trace}	6	0.828	0.488	0.238	[0.255, 1.426]
Synthetic	MRE	F^{out}	6	0.638	0.466	0.217	[0.147, 1.266]
		F^{trace}	6	0.630	0.182	0.033	[0.412, 0.861]

Table 12: Exploratory cross-dataset comparisons for Causal Estimation. No shared question IDs exist across datasets; Mann–Whitney U tests (p_{MW}) are unadjusted. "CI" records whether the reported 95% intervals for μ_A and μ_B overlap. The dagger marks $p < 0.05$ with non-overlapping intervals; the comparison is descriptive.

Metric	Type	Dataset A	μ_A	Dataset B	μ_B	$\Delta\mu$	CI	p_{MW}
MSA	F^{out}	Textbook	0.098	Real	0.236	0.138	✓	0.368
	F^{trace}	Textbook	0.538	Real	0.559	0.021	✓	1.000
MRE	F^{out}	Textbook	1.012	Real	0.697	0.314	✓	0.314
		Textbook	1.012	Synthetic	0.638	0.374	✓	0.148
		Real	0.697	Synthetic	0.638	0.059	✓	1.000
	F^{trace}	Textbook	1.109	Real	0.828	0.281	✓	0.394
		Textbook	1.109	Synthetic	0.630	0.478	×	$0.026^{\dagger}$
		Real	0.828	Synthetic	0.630	0.197	✓	0.589

Takeaway. Across these CE datasets, the MSA–MRE ordering remains stable despite numerical variation. Within this task family, a source-dataset estimate can therefore provide an initial structural ranking, followed by target-dataset re-estimation or held-out validation.

C.8 T2SQL Freedom Provenance

Table 13: Provenance and use of the T2SQL Freedom estimates. Canonical pre-synthesis means determine both the converter input and the matched structural coordinates in Table 7.

Estimate	Value	Use
Canonical F_{out}	0.686–1.071	Converter input and structural analysis
Canonical F_{trace}	0.147–1.249	Complementary process-space analysis

D PER-TASK FULL RESULTS

D.1 CAUSAL ESTIMATION

Table 14 reports the complete Causal Estimation comparison used in RQ1, including the scorer-controlled MSA–MRE reversal and deployment measurements.

Table 14: **Causal Estimation results.** Method choice is tightly coupled to MSA ($F{\approx}0.2$) and weakly coupled to MRE ($F{\approx}0.8$). The pipeline-faithful Workflow Compiler gains strongly on Textbook MSA but regresses on Synthetic MRE, while AdaSkill records the highest MSA and lowest MRE on all three datasets.

Dataset→	**Textbook**		**Synthetic**		**Real**		**Efficiency**	
Method↓	MSA↑	MRE↓	MSA↑	MRE↓	MSA↑	MRE↓	Time (s)↓	Cost ($)↓
CAIS (workflow ref.)	83.3	54.0	75.9	16.2	78.3	32.0	123	0.172
Base Agent	61.5	31.2	**100**	10.6	87.1	18.3	34	0.296
EvoSkill	71.8	25.4	**100**	7.1	93.5	21.6	66	0.270
Workflow Compiler	82.1	26.3	**100**	12.9	90.3	17.3	115	0.608
Tools Only	69.2	28.3	95.6	5.5	87.1	17.0	49	0.208
Knowledge Only	79.5	25.6	91.1	6.0	90.3	18.0	55	0.305
Pipeline Only	84.6	26.5	**100**	5.5	**96.8**	17.8	56	0.295
AdaSkill (Ours)	**89.7**	**23.1**	**100**	**4.3**	**96.8**	**15.9**	52	0.207

D.2 CAUSAL DISCOVERY: EXTENDED METRICS

Table 15 reports F1, precision, FPR, SHD, NHD, time, and cost across the four Causal Discovery datasets. Their canonical output-space estimates span $F^{\text{out}} = 0.291$–0.914. Results are mean $\pm$ std over 3 runs.

Table 15: **Causal Discovery metrics (mean $\pm$ SD, 3 runs).** SHD and NHD use sample SD (denominator $n-1$). P = Precision, FPR = False Positive Rate, SHD = Structural Hamming Distance; NHD = SHD/$[d(d-1)]$ for d nodes.

Dataset→	**AutoMPG**							**DWDClimate**						
Method↓	P↑	F1↑	FPR↓	SHD↓	NHD↓	Time↓	Cost↓	P↑	F1↑	FPR↓	SHD↓	NHD↓	Time↓	Cost↓
MATMCD (workflow ref.)	.71±.06	.74±.02	.12±.01	2.3±1.5	.12±.08	538.0±42.5	.69±.12	.60±.02	.66±.06	.10±.01	5.0±1.0	.17±.03	749.0±65.3	1.05±.15
Base Agent	.58±.07	.67±.05	.20±.05	3.0±1.0	.15±.05	165.7±35.9	2.00±.54	.55±.05	.56±.01	.13±.04	6.0±1.0	.20±.03	203.0±70.5	2.42±.26
EvoSkill	.62±.05	.70±.03	.17±.03	3.0±1.0	.15±.05	193.7±44.2	2.03±.63	.67±.05	.74±.03	.10±.02	4.0±1.0	.13±.03	205.0±33.8	2.48±.28
Workflow Compiler	.62±.07	.70±.03	.17±.03	3.0±1.0	.15±.05	52.3±8.0	.82±.08	.59±.03	.64±.08	.12±.03	4.7±.6	.16±.02	96.8±14.7	1.08±.02
Tools Only	.52±.03	.65±.02	.24±.03	3.7±.6	.18±.03	112.2±20.2	.43±.09	.52±.03	.54±.05	.13±.01	5.7±.6	.19±.02	123.1±64.9	.45±.11
Knowledge Only	.57±.03	.67±.02	.20±.02	3.0±1.0	.15±.05	22.9±1.8	.47±.08	.61±.05	.74±.06	.15±.02	4.0±1.0	.13±.03	30.2±2.7	.49±.09
Pipeline Only	.57±.02	.67±.03	.20±.01	3.0±1.0	.15±.05	22.6±1.0	.44±.07	.50±.07	.61±.10	.19±.02	6.0±1.7	.20±.06	32.0±5.7	.42±.08
AdaSkill (Ours)	.56±.08	.65±.05	.22±.06	3.3±1.2	.17±.06	112.0±1.2	1.99±.31	.65±.07	.70±.05	.11±.04	4.0±1.0	.13±.03	169.3±51.8	2.62±1.57
Dataset→	**Sachs**							**Child**						
Method↓	P↑	F1↑	FPR↓	SHD↓	NHD↓	Time↓	Cost↓	P↑	F1↑	FPR↓	SHD↓	NHD↓	Time↓	Cost↓
MATMCD (workflow ref.)	.58±.10	.56±.11	.06±.01	13.3±12.7	.12±.12	2747.0±185.5	3.82±.45	.65±.06	.65±.05	.02±.01	14.3±10.2	.04±.03	3774.0±256.2	6.49±.72
Base Agent	.86±.03	.87±.05	.03±.01	4.0±1.0	.04±.01	153.6±2.9	1.78±.53	.77±.17	.80±.17	.02±.01	9.3±9.1	.02±.02	201.8±51.4	2.66±.52
EvoSkill	.86±.08	.87±.08	.03±.02	4.0±1.7	.04±.02	154.0±39.9	1.39±.13	.88±.13	.90±.10	.01±.01	5.0±6.1	.01±.02	457.2±133.2	4.83±.72
Workflow Compiler	.75±.11	.67±.26	.03±.01	7.7±5.7	.07±.05	198.0±109.1	1.95±1.11	.86±.05	.86±.05	.01±.01	6.7±3.5	.02±.01	328.9±131.5	2.48±.65
Tools Only	.86±.08	.89±.08	.03±.02	4.0±1.7	.04±.02	92.9±62.9	.29±.05	.90±.05	.85±.03	.01±.01	5.7±1.2	.01±.00	115.6±23.2	.54±.15
Knowledge Only	.88±.05	.89±.05	.02±.01	3.3±1.5	.03±.01	54.9±9.4	.41±.07	.81±.08	.79±.07	.01±.01	9.3±4.0	.02±.01	161.4±103.2	.54±.11
Pipeline Only	.94±.05	.95±.05	.01±.01	1.7±2.1	.02±.02	75.9±65.1	.45±.08	.79±.05	.81±.04	.02±.01	9.0±2.6	.02±.01	55.0±18.2	.77±.14
AdaSkill (Ours)	.91±.07	.92±.06	.02±.01	2.3±2.1	.02±.02	301.8±212.5	2.82±1.11	.96±.06	.95±.07	.00±.00	1.7±2.9	.00±.01	170.6±12.7	2.55±.62

D.3 FEATURE ENGINEERING

Feature Engineering operates in a high-freedom regime: the canonical output-space estimates are $F^{\text{out}} = 1.037$–1.337. Table 16 reports AUC, selected feature counts, latency, and cost. AUC differences are small, while removing the evolutionary loop improves deployment efficiency.

D.4 TEXT-TO-SQL

Table 17 reports the complete execution and schema-declaration results for BIRD-147 and Spider-120, together with task-level latency and cost.

Table 16: **Feature Engineering results (AUC, mean ± std).** In these high-F settings, AdaSkill achieves the best mean AUC on Taobao and matches the source FELA workflow on Dia at substantially lower reported cost. The FELA workflow cost is estimated from its recorded operating range. EvoSkill's domain-average latency and cost are reported in Table 1; – denotes unavailable per-dataset details. High-F feature engineering favors minimal structure.

	Taobao					Dia				
Method	AUC	Max	#F	Time (s)	Cost ($)	AUC	Max	#F	Time (s)	Cost ($)
FELA (workflow ref.)†	.653	.653	N/A	32400	11.46	**.812**	.812	N/A	43200	15.80
Base Agent	.660$_{\pm.005}$	.667	59	2235	3.75	.804$_{\pm.006}$	.811	76	3042	6.48
EvoSkill	.664	–	–	–	–	.806	–	–	–	–
w/ Workflow Compiler	.658$_{\pm.008}$	.666	48	2682	3.09	.798$_{\pm.002}$	.800	18	2466	2.47$_{\pm2.18}$
AdaSkill (Ours)	**.670**$_{\pm.002}$	**.672**	46	1972	3.48$_{\pm1.72}$	**.812**$_{\pm.001}$	**.813**	55	2301	5.05$_{\pm1.57}$

Table 17: **Text-to-SQL results.** SRR, NSR, NSP, and NSF score the explicit `selected_columns` declaration; EX independently scores the executed result. Their divergence on Spider records incomplete declarations alongside executable SQL. AdaSkill improves all eight declaration metrics over the Base Agent and raises EX on both datasets.

Dataset→	BIRD-147					Spider-120					Efficiency	
Method↓	SRR↑	NSR↑	NSP↑	NSF↑	EX↑	SRR↑	NSR↑	NSP↑	NSF↑	EX↑	Time↓	Cost↓
APEX (workflow ref.)	**93.9**	**98.3**	47.0	60.8	**69.4**	**88.3**	**97.1**	36.5	47.0	**67.5**	361	0.72
Base Agent	71.8	91.8	90.0	90.5	59.2	0.8	8.4	21.7	11.4	60.0	75	0.39
EvoSkill	76.2	92.5	94.7	93.0	66.0	2.5	15.2	32.5	20.8	62.5	97	0.48
Workflow Compiler	79.6	93.6	94.8	93.7	67.4	1.7	10.3	22.6	13.4	57.5	143	0.59
Tools Only	75.5	92.5	94.7	92.9	66.0	3.3	20.5	45.0	28.5	63.3	104	0.37
Knowledge Only	74.2	92.5	**94.9**	93.1	66.7	5.0	36.4	68.3	45.3	64.2	93	0.51
Pipeline Only	68.7	90.7	94.2	91.7	64.6	1.7	12.0	28.3	16.2	60.8	124	0.88
AdaSkill (Ours)	80.8	94.2	**94.9**	**94.1**	67.4	7.5	51.3	**85.8**	**60.5**	66.7	101	0.45

E Efficiency and Cost

E.1 Performance and Cost Pareto Analysis

Figure 14 compares dataset-mean performance and reported deployment cost on the same basis as Table 1, without an additional cache adjustment. AdaSkill lies on the non-dominated frontier among the displayed methods in all six panels. The T2SQL EX panel shows a trade-off: the workflow reference has 1.4 points higher mean EX and 60% higher measured cost. The remaining panels show either performance dominance by AdaSkill or a frontier trade-off with a cheaper baseline.

E.2 Per-Dataset Performance and Efficiency

Figure 15 provides per-dataset AdaSkill performance across all 11 datasets (Row 1). Rows 2–3 report task-average efficiency for T2SQL and CE, for which the recorded rows contain aggregate time and cost, and per-dataset efficiency for CD and FE.

F Workflow-to-Skill Conversion Case Study: Causal Estimation

We trace the transformation from the original CAIS workflow (Verma et al., 2025) to the directly deployed AdaSkill, illustrating the component boundary and the metric-conditioned structural decision. The architecture decision is conditioned on CE-MSA.

Adaptive Distillation and Architecture Transformation

Figure 16 contrasts the original CAIS four-stage, eight-micro-tool workflow with the resulting skill-guided tool-using agent.

What happens to the retained pipeline. Under the functional boundary in Appendix C.2, the original `method_selector` is pipeline, not knowledge: it maps observed study properties to an estimator choice. CE-MSA falls on the retain route, so AdaSkill preserves this metric-qualified

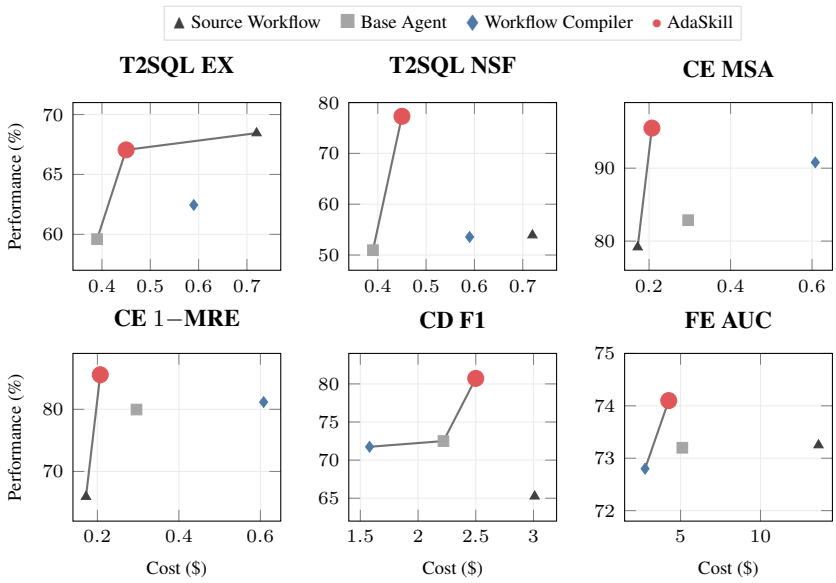


Figure 14: Performance and reported deployment cost on the Table 1 basis; performance is averaged equally across datasets. No additional cache adjustment is applied. Solid segments connect non-dominated displayed methods. AdaSkill lies on this frontier in every panel; T2SQL EX retains a trade-off between accuracy and cost relative to the workflow reference.

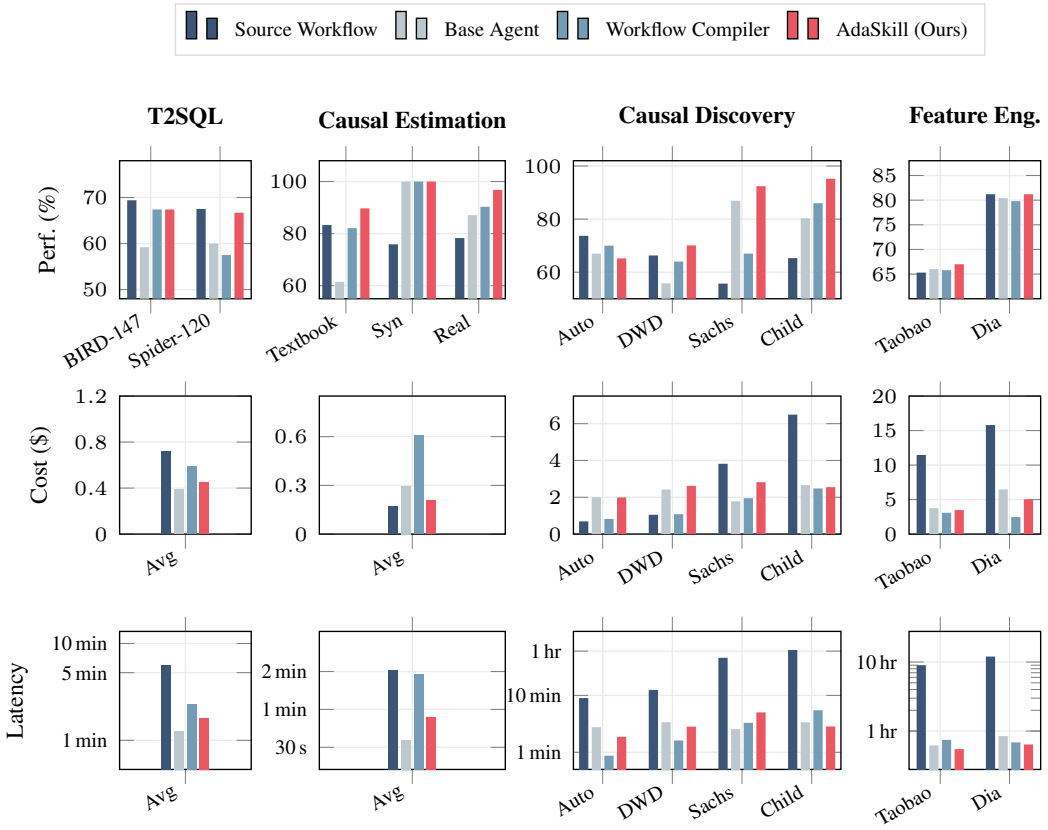


Figure 15: **Final performance and efficiency breakdown. Row 1**: selected final-skill performance per dataset (EX% / MSA% / F1×100 / AUC×100). **Rows 2–3**: selected-final efficiency; T2SQL/CE show the task averages reported in Tables 17 and 14, while CD/FE use the per-dataset latency and cost in Tables 15 and 16, on the Table 1 accounting basis. No additional cache adjustment is applied. †FELA cost is estimated: Taobao $7.5–$15, Dia $10–$22 per run.

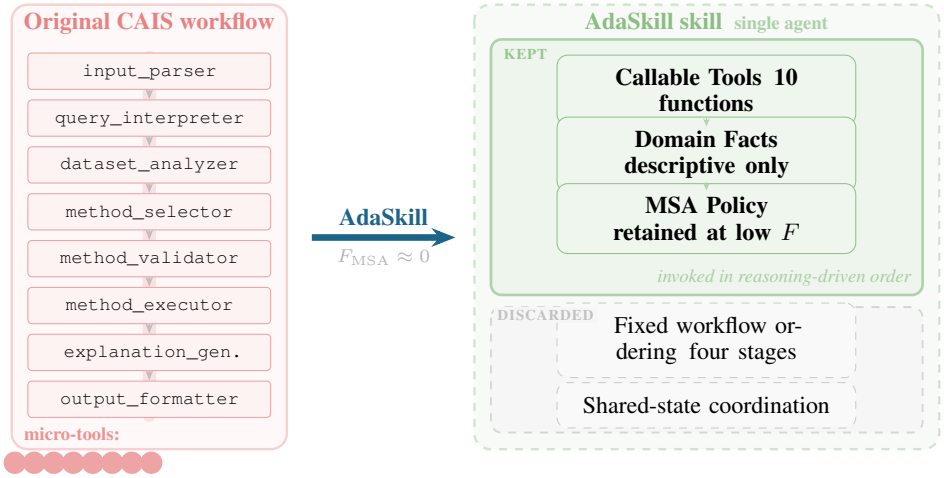


Figure 16: **AdaSkill architecture transformation** for CE ($F_{\text{MSA}} \approx 0$). **Left**: the published CAIS workflow comprises four stages and eight shared-state micro-tools; method selection and validation are deterministic, and only a subset of the other tools use LLM assistance. **Right**: the adaptive skill retains callable tools and descriptive domain facts, expresses the low-F method-selection policy as a metric-scoped pipeline, and discards the source workflow's global ordering and shared-state orchestration. The diagram compares workflow structure, not model-call counts.

policy while removing the source workflow's global execution order. The source encodes the policy as hard-coded Python:

Original CAIS: rule_based_select_method (excerpt)

```
if is_rct:
    if instrument_var and instrument_var != treatment:
        return {"selected_method": "instrumental_variable", ...}
    if covariates:
        return {"selected_method": "linear_regression", ...}
    else:
        return {"selected_method": "diff_in_means", ...}
```

The adaptive skill converts the same retained pipeline to metric-scoped prose in `SKILL.md`:

AdaSkill SKILL.md: retained low-F method-selection pipeline

```
### Step 1 - RCT check (do this first)
If description contains: randomized / RCT / randomly assigned -> is_rct = True
- When is_rct = True: matching, PSM, and IPW are FORBIDDEN.
  OLS with covariates is the correct choice.
  Random assignment already balances covariates --
  applying matching on top produces an incorrect selected_method.
- If an instrument for non-compliance exists -> use IV (encouragement design).
```

These are verbatim benchmark-facing rules, not universal causal-inference prescriptions: randomization does not guarantee finite-sample covariate balance or prohibit every valid weighting estimator. The imperative and prohibitive clauses are deliberately classified as pipeline. Their presence is evidence that the low-F route retained method-selection structure, not that pipeline was hidden inside knowledge. A discard-route module removes the entire block; descriptive facts about randomization and estimator assumptions remain available without prescribing an action.

F.1 Exploratory Execution Case: Construction and Audit

Figure 10 is a new qualitative illustration on the electrification case, not an additional held-out benchmark aggregate. The case had already been inspected before these candidates were constructed. Each condition has one completed Sonnet 4.6 execution; there was no retry, seed search, or result-dependent candidate revision in this three-arm comparison. Earlier exploratory runs with a manually reconstructed Selective prompt are excluded: that prompt was not the released AdaSkill.

Candidate provenance and intervention boundary. The AdaSkill text is byte-identical to the released CE skill and to the supplied source snapshot (SHA256 prefix `744bc4e5f7999df2`). Minimal deletes skill-text pipeline constraints while retaining separable descriptive content, tool signatures, and the output interface; mixed clauses have a line-level disposition audit. Full appends control rules mapped from the accompanying CAIS source snapshot: stage order, candidate priority, validation, and backtracking within the released tools' supported methods. Unsupported methods and native orchestration are not added. Thus Full is a source-mapped skill conversion, not a native MAS execution. Original tool code, including advisory outputs, remains identical across arms; Minimal is not a wholly guidance-free runtime. The three conditions share the CSV, tool resources, model, dependency versions, and per-run budget. Post-execution hashes confirm that these resources and the frozen skill files were unchanged. No F or threshold is re-estimated.

Numerical evidence and estimands. The legacy reference is method IV with value 4509.4; reference-based deviation is $|\widehat{\theta} - 4509.4|/4509.4$. Figure 10 instead shows source text, actual calls, and selected returned fields. Independent linear-algebra replay reproduces AdaSkill's IV estimate 4049.358305, Full's IV estimate 1469.962126, Minimal's raw RDD jump -770.386191, and Full's intermediate RDD jump 123.810369. AdaSkill's covariates are age, religion, gender, and caste; Full's final IV adds distance. Both IV fits use 685 complete cases. The Full RDD bandwidth is 15 m (463 observations); Minimal uses 20 m (685 observations). The shared RDD tool uses the supplied covariates only for complete-case filtering, not as regression predictors. Consequently neither RDD call is described as covariate-adjusted.

Minimal flips the RDD outcome jump's sign and labels it the effect of actual connection. Yet connection rates are approximately 0.599 among eligible and 0.020 among ineligible households. An eligibility-side outcome contrast is not a treatment effect under incomplete compliance without the corresponding first-stage normalization and identification assumptions. Moreover, no data lie strictly between 35 and 45 m: a cutoff estimate requires extrapolation across this gap, not an automatically valid local RD comparison.

Observed compliance and interpretation limits. Full follows the restored RDD priority and source gap heuristic, declares validation passed, and executes RDD. It then switches to IV after seeing a small, imprecise estimate and noting the support gap. This is not the prescribed pre-execution validation-failure backtrack. Its final covariate choice also differs from AdaSkill; this single run does not isolate which clause caused that choice or establish that adding distance is scientifically incorrect. The case illustrates observed execution differences, not perfect compliance with the source program or a controlled clause-level causal effect.

Reference proximity is not causal validity. The legacy reference's full estimation specification remains unresolved: replay with the metadata-listed covariates gives approximately 4053.61, not 4509.4. In addition, the models' verbal claims that exclusion and exogeneity hold are stronger than the evidence: instrument relevance testing alone establishes neither. An IV/LATE interpretation requires further identification assumptions; the output field `ate` does not establish a population ATE. All three outputs are retained regardless of reference deviation. The case is descriptive and supports no population-level significance claim.

Additional routing, deployment and sensitivity detail. Pre-synthesis inputs are CE MSA 0.167, CE MRE 0.782, CD F1 0.602, T2SQL EX 0.686, T2SQL NSF 0.816, unidentifiable T2SQL SRR, and FE AUC 1.037. CE and CD use pre-specified means over diagnostic datasets; T2SQL and FE use their designated conversion benchmarks. We first check source qualification and target interfaces, learn Retain versus Discard from matched development effects, then freeze and execute on held-out regimes. Richer response-free states can tighten the representation remainder; residual within-behavior quality changes are not assumed away.

The \$6.12 budget covers one CE 6×6 diagnostic execution unit, not total diagnostic overhead or construction. Main-benchmark CE execution costs are \$0.296 (Base), \$0.223 (Full), \$0.211 (Discard), and \$0.207 (AdaSkill) per query. Budget, backbone, and uncertainty analyses are reported in Section 4.6; these benchmark aggregates are distinct from the exploratory one-case comparison.