DATA VALUE DEPENDS ON THE FULL TRAINING STATE

Anonymous authors
Paper under double-blind review

ABSTRACT

A post-training pipeline has to choose its next dataset before it can read any outcome, so it decides from a score computed ahead of the run. Every such score returns one verdict per dataset, read from the data alone or from the data together with the current parameters. We show that the usefulness of a dataset is a property of the data together with the state that reads it, and that this state carries more than the weights: it carries the optimizer memory and the step clock as well. Two states can therefore share a verdict on a dataset whose effects at those states have opposite signs, and the shared verdict is wrong at one of them. We call the effect at the receiving state *causal trace value* (CTV) and measure it directly, by restoring the saved state and continuing once on the candidate and once on the incumbent at equal tokens and steps, so that the paired difference in held-out accuracy credits the choice of data alone. The measurement confirms itself before it reports, replaying one continuation and returning no verdict when the replay disagrees. Any rule whose verdict two states share pays a summed regret of at least the smaller of the two effects on a dataset whose effects at those states have opposite signs, so the price of a shared verdict is a quantity. On a frozen benchmark of text-to-SQL decisions, measurement achieves positive mean validation utility over twelve training configurations at a matched admission count, admits a fraction of the candidates that every rival policy must match, and trains continuations whose price we report beside what they buy.

1 INTRODUCTION

At a saved checkpoint with a fixed training budget, a team must choose which dataset to train on next before any result comes back. Current practice supports this choice with a score. Attribution methods credit model behaviour to training examples (Koh & Liang, 2017; Ghorbani & Zou, 2019; Ilyas et al., 2022; Park et al., 2023), and selection scores compare a candidate with a small held-out set through gradient alignment, distribution matching, or a learned selection reward (Pruthi et al., 2020; Xia et al., 2024; Sun et al., 2026). Each reduces a dataset to one number, computed from the data alone or from the data together with the current weights, and that number is taken as the value of the dataset.

We take this assumption as the object of study. The effect of a dataset depends on the state that receives it, and that state holds more than the weights (Hu et al., 2026b; Dang et al., 2026). Distillation studies report that stronger teachers can produce worse students, with a gap governed by student capacity and data–model compatibility, and that high-reward traces can impair small-model reasoning (Bansal et al., 2025; Huang et al., 2026; Qian et al., 2026). The optimizer also carries state, namely running averages of past gradients and a count of the steps already taken (Kingma & Ba, 2015; Loshchilov & Hutter, 2019); finite-horizon effects, delayed transport of optimizer state, stored optimizer state, and the gap between an adaptive optimizer's applied direction and the raw gradient all shape the next update (Liu & Li, 2026; Guo, 2026; Xu, 2026; Xiao et al., 2026). Two checkpoints with identical weights can therefore respond to the same dataset in opposite directions, and a rule that cannot distinguish two such states has the same admission probability at both.

Choosing between a candidate and the incumbent from a measured effect is a treatment choice under uncertainty, and the regret of a decision rule can be bounded against the alternative it forgoes (Manski, 2004; Hirano & Porter, 2009; Kitagawa & Tetenov, 2018; Montiel Olea et al., 2026). A data-level score can replace a measurement at the state exactly when the effect of a dataset is the same at every receiving state, that is, when every interaction between datasets and states vanishes (Manski & Tetenov, 2007; Stoye, 2009; Athey & Wager, 2021). When a dataset's effect changes sign between

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

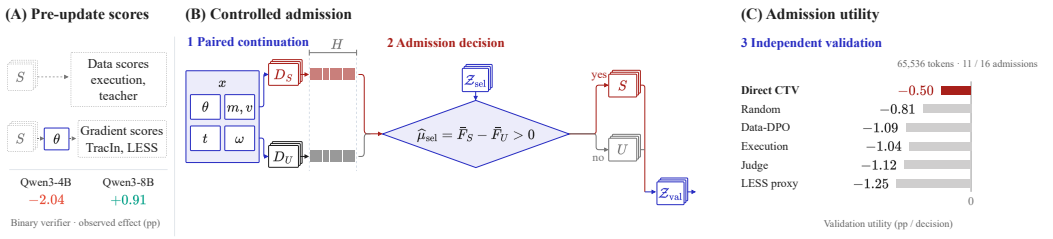

Figure 1: Scoring a candidate update against measuring it. (A) A score reads the data or the receiving parameters. (B) The measurement continues once on the candidate and once on the incumbent from one saved state, so the paired difference credits the choice of data. (C) Validation utility at the count the measurement chooses, on sixteen cross-state decisions at 65,536 tokens.

two states and the rule's input is identically distributed at those states, the rule's summed regret is at least the smaller effect magnitude. The same analysis bounds the number of repeated runs a measurement needs, so the seed budget is a calculable quantity.

We therefore measure the effect at the receiving state and call the measurement *causal trace value* (CTV). CTV restores the saved state exactly, including both optimizer moments, the step clock and the random state, then continues training once on the candidate dataset and once on the incumbent with identical token counts and numbers of optimizer steps. The two continuations share every input except the dataset, so the difference in held-out accuracy is attributable to the choice of data, and the candidate is admitted when that difference is positive across repeated runs. Because the comparison requires both continuations to start from the same state, CTV repeats one continuation from the restored state and withholds a verdict when the repeat disagrees with the first. The cost is the continuations themselves, which the scores CTV replaces never train.

We evaluate CTV on a frozen benchmark of 144 text-to-SQL decisions, a setting in which small curated sets, semantics-aware synthesis, and agentic reinforcement learning each change what a trace set teaches (Ma et al., 2026; Gao et al., 2026; Kakkar et al., 2026). We compare policies at CTV's admission count of 42 candidates. Averaged over all 144 decisions, validation utility relative to retaining the incumbent is $+0.17$ percentage points for CTV, -0.48 in expectation for uniform random selection, and $+0.02$ for the strongest evaluated scoring baseline. We report what the measurement costs beside what it buys, and we report where the choice of data stops paying at all.

2 Related Work

Scores for data selection. Proxy-model scores (Yu et al., 2024; Mindermann et al., 2022), and later work that widens the inputs of a per-example score or lowers its cost (Lin et al., 2026; Dagli et al., 2026; Chen et al., 2026b; Ding et al., 2026), all return a number per dataset or per example, which is the object this paper studies. Source reweighting during training pursues the same objective by adjusting mixture weights rather than admitting datasets (Hu et al., 2026a; Wang et al., 2026c;b; Chen et al., 2026a). Audits report that influence-selected examples help little until their responses are rewritten, that values shift with validation structure, and that attribution can track step length in place of the update it estimates (Luo et al., 2026b; Shen et al., 2026; Wang et al., 2026a).

Value that moves with the state. Prior work observes that a data value changes with the state that receives it: across iterations and developmental stages of one trajectory (Wang et al., 2024; Lee et al., 2025), across reference checkpoints (Nishida et al., 2026), across optimizers (Ding et al., 2026), and across later training stages (Yang et al., 2026). These findings describe how a value moves between states. Corollary 1 instead bounds the regret of every decision rule whose input is identically distributed at two states with opposite-sign effects.

The receiving student. Iterative synthetic tuning avoids collapse by choosing what each generation trains on (Alemohammad et al., 2026; Luo et al., 2026a). LARK and DIVE select data by the receiving model's learnability and self-influence, task learnability and online-verified prompt selection track the current policy, and self-evolving agents check each new skill before committing it (Yu et al.,

2026; Nyakairu & Liu, 2026; Zheng et al., 2026). These methods share the concern with student compatibility that motivates our measurement; CTV addresses it with a paired continuation from the saved state, including the optimizer moments.

What a score can decide. A better score need not yield a better decision: Data Shapley and similar semivalues can select subsets no better than random, and NASH responds by splitting validation accuracy into Shapley-informative components and aggregating them non-linearly (Tian et al., 2026). NASH improves which dataset a score ranks, and Corollary 1 bounds the regret in the cases where no single ranking serves two receiving states.

3 PRELIMINARIES

A post-training run resumes from a saved *receiving state*

$$x = (\theta, m, v, t, \omega), \tag{1}$$

which collects the trainable parameters θ, the first and second AdamW moments m and v, the per-parameter step clock t, and the random state ω of the training framework (Loshchilov & Hutter, 2019). The state trains through low-rank adapters, so θ names the saved factors and m and v name the optimizer state that trains them (Hu et al., 2022).

The pipeline holds an *incumbent* dataset U and a shortlist of *candidate* datasets $S_1, \ldots, S_M$. Packing turns each dataset into a pack of training tokens with a fixed token count and a fixed number of optimizer steps, written D_U and D_S for the incumbent and one candidate. We write $B(x, D)$ for the state reached by continuing training from x on pack D, $F(y)$ for the mean execution accuracy of the model at state y on the validation questions, and $\hat{F}(y)$ for the same accuracy on a disjoint selection split. The candidate is read against the incumbent through the paired difference

$$\hat{\tau}(x; S) = \hat{F}(B(x, D_S)) - \hat{F}(B(x, D_U)), \tag{2}$$

which we call the *causal trace value* of S at x and report in accuracy points, with $\tau(x; S)$ the same difference under F. Both continuations start from x, share their token count and their step count, and differ in the pack alone, so the difference in Equation 2 credits the choice of data (Rubin, 1974; Ilyas et al., 2022).

A *policy* reads whatever it is allowed to read and returns an accept or a retain for each candidate. Its *realized utility* is the F-difference against the incumbent for an accept and zero for a retain, so every policy is scored on questions it did not read during the decision, and an accept whose realized utility is negative is a *harmful admission*. The policies this paper compares read the dataset alone, read the dataset together with the current parameters, or train a small proxy that stands in for the receiving model.

4 WHY ONE VERDICT PER DATASET FAILS

This section establishes the problem the measurement solves. We first record two measurements that show what the receiving state contains. We then state the exact condition under which a data-level score can replace a measurement at the state, and price the verdicts taken when that condition fails.

4.1 WHAT THE RECEIVING STATE CONTAINS

The first measurement concerns who does the reading. A trace set is usually valued by how far it raises the accuracy of the model that produced it, and that valuation does not transfer to the model that trains on it. Among four trace sets drawn from database-disjoint sources, the set with a 13.5-point gain at its teacher moved the student's accuracy by -5.2 points, and the student responses to the four sets spread over 12.5 points (Figure 2A). The two readers here are two sizes of one model family, so the reversal is not an artefact of comparing unrelated systems.

The second measurement concerns what the state holds. Restoring the parameters alone resets the optimizer memory, which changes the bias correction and the effective step size of every coordinate. Feeding one student the same tokens in two orders gave an order contrast of -1.30 points when

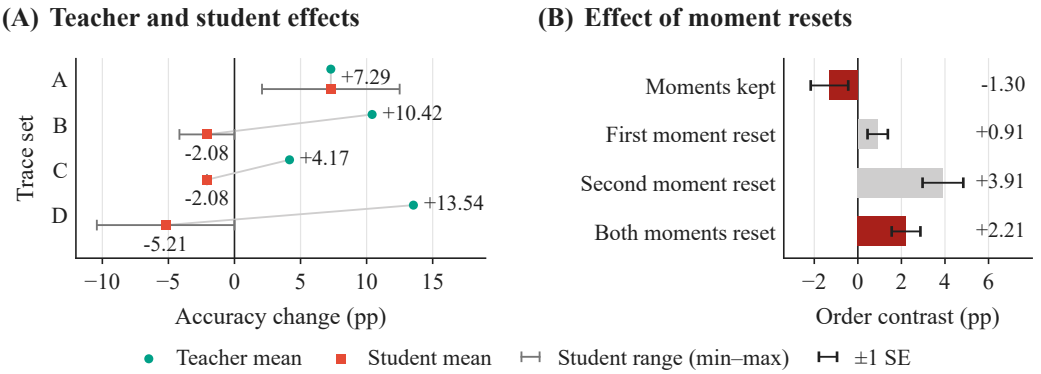

Figure 2: Two measurements of the receiving state. (A) Teacher accuracy gain and student accuracy change for four trace sets. (B) Order contrast under AdamW moment resets.

Table 1: Policy inputs and whether decisions can change with the receiving state. Appendix D.1 lists the policies inside each class and cites the method each one reduces.

Input class	Policies	Decision at two states
Dataset only	5	shared by construction
Dataset through θ	10	can differ
Proxy model	4	shared or partial
$(\theta, m, v, t, \omega)$	2	measured at each state

both moments were carried across the interruption, and 3.52 points when both moments were reset (Figure 2B). A checkpoint that stores parameters alone therefore leaves the effect of a dataset undetermined.

4.2 When a score can stand in for the measurement

Let $\mathcal{X}$ be a finite set of receiving states, $\mathcal{A}$ a finite set of datasets, and $F(x, a)$ the validation accuracy after continuing x on dataset a. Fix anchors $x_0 \in \mathcal{X}$ and $a_0 \in \mathcal{A}$ and define the anchored interaction

$$\Delta(x, a) = F(x, a) - F(x, a_0) - F(x_0, a) + F(x_0, a_0), \tag{3}$$

which measures how far the effect of a against a_0 moves when the receiving state moves from x_0 to x.

Theorem 1 (Transferable dataset values)**.** *Every anchored interaction vanishes exactly when $F(x, a) = b(x) + g(a)$ for some functions b and g. The components $b(x) = F(x, a_0)$ and $g(a) = F(x_0, a) - F(x_0, a_0)$ are then the unique pair with $g(a_0) = 0$. When every $|\Delta(x, a)| \le \varepsilon$, those components reconstruct the table to within ε.*

The function g is the part of a dataset's value that carries across states, and b is the part each state keeps. Under the additive form the effect $\tau(x; S) = g(S) - g(U)$ is one number at every state, and a data-level score can at best recover it. A dataset whose effects τ_1 and τ_2 at states x_1 and x_2 have opposite signs carries interaction of size $|\tau_1| + |\tau_2|$ between those states, so no such number exists for it. The next statement prices the consequence.

4.3 The price of a shared verdict

Fix a dataset with effects τ_1 and τ_2 at states x_1 and x_2. Admitting at state i realizes τ_i and retaining realizes zero, so the regret of a retention is τ_i^+ and the regret of an admission is $(-\tau_i)^+$. Call a rule *state-free* when its input is identically distributed at the two states, so that it admits with one probability $p \in [0, 1]$ at both.

Algorithm 1 Causal trace value: admission by paired continuation from the saved receiving state.

Require: receiving state $x = (\theta, m, v, t, \omega)$; token- and step-matched packs D_U, D_S; updater seeds $1, \ldots, n$; selection and validation splits Q, V.

for $j = 1, \ldots, n$ **do**
  Set ω by seed j and save the state
  for $D \in \{D_U, D_S\}$ **do**
    Restore the state; $x_D^{(j)} \leftarrow B(x, D)$
  end for
  $d_j \leftarrow \hat{F}(x_{D_S}^{(j)}) - \hat{F}(x_{D_U}^{(j)})$
end for
Restore the state at seed 1 and replay $B(x, D_U)$
if the replay outcomes differ from $x_{D_U}^{(1)}$ **then**
  return STOP
end if
$\bar{d} \leftarrow \frac{1}{n} \sum_j d_j$
return admit S if $\bar{d} > 0$, else retain U

Corollary 1 (Sign-flip price)**.** *Suppose $\tau_1 > 0 > \tau_2$. Every state-free rule has summed expected regret $(1-p)|\tau_1| + p|\tau_2| \geq \min(|\tau_1|, |\tau_2|)$ over the two states, and the case $\tau_1 < 0 < \tau_2$ exchanges the two effects. A deterministic state-free rule has regret at least $\min(|\tau_1|, |\tau_2|)$ at one of the states, and every randomized one has worst-state expected regret $\max\{(1-p)|\tau_1|,\, p|\tau_2|\} \geq |\tau_1||\tau_2|/(|\tau_1| + |\tau_2|)$, with equality at $p = |\tau_1|/(|\tau_1| + |\tau_2|)$.*

The proof is the degenerate case of Le Cam's two-point method, with identically distributed input under both states. Table 1 sorts the implemented policies by what they read. Admit-all, teacher accuracy, execution correctness, near-duplicate removal, and convex dataset valuation read the dataset alone and share one verdict by construction, so the corollary covers them on every dataset whose sign changes. Perplexity, the judge on the receiving model, TracIn, LESS, TSDS, Data-DPO, and PRISM read each state's parameters and can change their verdict, and Appendix B.3 reports their registered decisions at both states.

5 Measuring the Effect at the Receiving State

The measurement restores the saved state, continues twice at an equal budget, and reads one paired difference. Each repeated run sets the random state, and the estimate of the effect in Equation 2 is the mean of the paired differences over the runs. The candidate is admitted when that mean is positive on the selection questions; otherwise the incumbent is retained. Separate validation questions evaluate the realized utility. Both continuations share every input apart from the pack, which is what lets the difference credit the choice of data.

Because the comparison requires both runs to start from the same state, the measurement checks itself before it reports. It repeats one continuation from the restored state on the same pack, and deterministic execution must return identical per-question outcomes. A measurement whose repeat disagrees returns no verdict at all, which gives the instrument a validity test of its own. Appendix C.7 states the conditions of the check, and Appendix D reports a serving configuration that fails it.

A pipeline that holds several packs built by different selection rules at one budget, together with the incumbent pack, can apply the same instrument to the whole menu. Writing the incumbent as D_0, the menu rule continues on every pack from the same state and trains on

$$j^\star \in \arg\max_{0 \leq j \leq M} \hat{F}(B(x, D_j))\,, \tag{4}$$

taking $j^\star = 0$ on ties. The menu assigns no score to any example, and its price is one continuation per pack together with one selection read per pack.

The measurement pays an error of its own, which comes from updater noise, and the size of that error sets the number of repeated runs. Let the paired differences be $d_1, \ldots, d_n$ with $d_j = \tau + e_j$, where τ

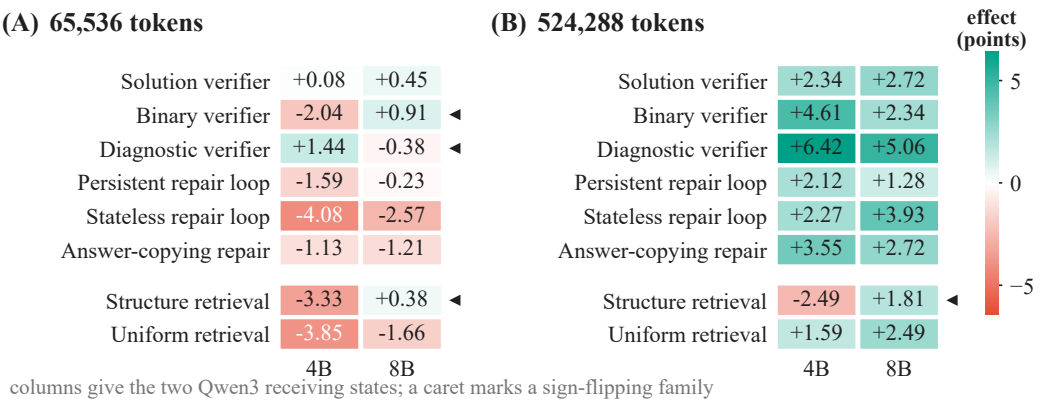

Figure 3: Candidate-minus-incumbent validation effects by producer family and receiving state, in percentage points. (A) 65,536 tokens. (B) 524,288 tokens.

is the effect on the fixed selection questions and the e_j are independent, mean-zero and sub-Gaussian with variance proxy σ^2, given a shared saved state and random stream within each run.

Proposition 1 (Seed complexity)**.** *The rule that admits when $\bar{d} > 0$ errs in sign with probability at most $\exp(-n\tau^2/2\sigma^2)$. Hence, for a floor $\tau_{\min} > 0$, $n \geq 2\sigma^2 \log(1/\delta)/\tau_{\min}^2$ runs decide the sign of every effect with $|\tau| \geq \tau_{\min}$ with error at most δ, and the expected regret is at most $|\tau| \exp(-n\tau^2/2\sigma^2) \leq \sigma/\sqrt{\mathrm{e}\, n}$ for every τ.*

The sub-Gaussian tail bound gives the first statement, and maximizing $t \exp(-nt^2/2\sigma^2)$ over $t > 0$ gives the uniform one. Appendix C.5 combines a one-run measurement with a pooled estimate from the other candidates and reports its utility at shorter seed budgets.

6 Experiments

Every outcome is BIRD text-to-SQL execution accuracy, where a generated query succeeds when execution returns the gold result (Li et al., 2023; Zhong et al., 2020). Every policy shares the same receiving checkpoint, tokenizer, SQL executor, data splits and updater seeds; only the rule that chooses the dataset differs. We organise the experiments around four questions. **Q1** asks whether measurement ranks candidate datasets above the data-level and parameter-level scores when the policies admit the same number of candidates (§6.1). **Q2** asks how each score tracks the measured candidate effect (§6.2). **Q3** asks where the choice of data stops mattering (§6.3). **Q4** asks what the measurement costs, and what a shorter schedule recovers (§6.4).

The frozen benchmark crosses twelve training configurations with two question-role directions and six evaluation criteria to yield 144 decision views. Each configuration fixes a trace family, receiving-model scale, adapter rank, and token budget, and is repeated at four updater seeds. The views reuse the trained models; selection and validation questions are paired within eleven databases. Configurations, databases, and seeds retain their separate resampling roles (Appendices B.2 and C.2).

The cross-state study separately evaluates eight producer families at Qwen3-4B and Qwen3-8B, giving sixteen decisions per dose. Each continuation uses 32 optimizer steps at 65,536 or 524,288 tokens. Policies read 179 selection questions and realized utility reads 441 validation questions on disjoint databases, averaged over three updater seeds. One incumbent continuation is shared across candidates at each state, dose, and seed, giving 108 continuations in total. Appendix B.3 specifies the registration and paired bootstrap.

6.1 What the Measurement Buys

Figure 4 compares policies at a common admission count, whereas Table 2 reports each score's agreement with measured effects. The positive selection-set effect rule determines the admission count without validation data; each policy is then evaluated on its own top-ranked candidates at

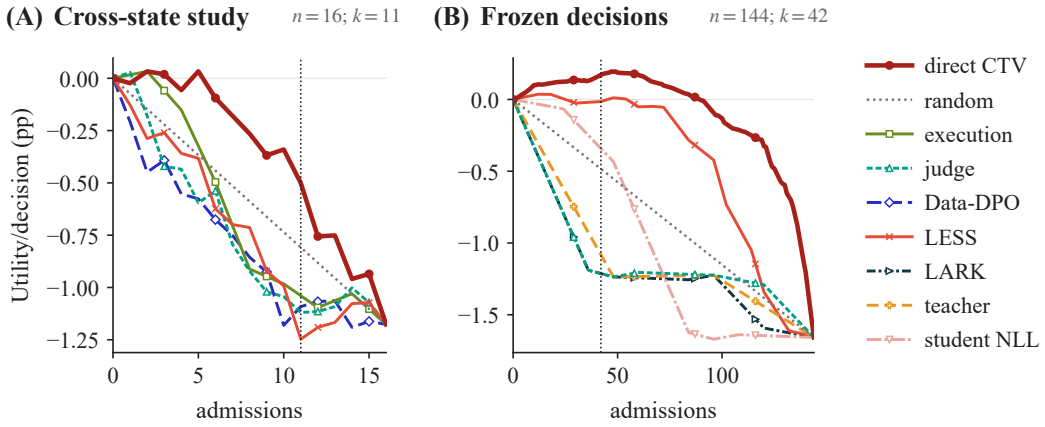

Figure 4: Realized utility by admission count and policy; the dotted line marks the Direct CTV count. (A) Sixteen cross-state decisions at 65,536 tokens. (B) 144 frozen decisions.

that count. Measurement admits 42 of the 144 frozen decision views and realizes $+0.17$ points per decision. Figure 4B compares rankings at that count; Table 6 reports utility, harm, and cost under each policy's own admission rule. Of the 42 admissions, 30 gain on validation, 11 lose, and one ties the incumbent. Random selection at the same count realizes -0.48 points per decision. On the small-dose cross-state study in panel A, measurement has utility -0.50 points relative to the incumbent: a smaller observed loss than the compared scores, with positive one-sided lower bounds against the judge and LESS-style proxy but bounds including zero against execution correctness, Data-DPO, and random selection.

Resampling the twelve training configurations gives a positive mean in 0.99995 of draws. Database-cluster resampling recomputes the policies on resampled questions and gives $+0.05$ points in $[-0.15, +0.21]$ (Appendix C.2). The gain resolves at the decision and configuration levels and remains open at the database level. These intervals describe that resampling analysis, not the curves in Figure 4B.

A same-pool follow-up asks whether family-level measurement also helps choose individual traces. On previously inspected validation data, CTV family selection exceeds a BIRD adaptation of the official LESS pipeline by $+5.06$ points, with one-sided lower bound $+1.23$, but its gap to uniform selection is -0.53 and its utility relative to the incumbent is -2.19 (Appendix D.1). The ranking gain over LESS therefore does not establish a beneficial update or a gain over random selection.

A separate post hoc study calibrates thresholds on 161 development questions. The calibrated thresholds lower utility from -0.50 to -0.73 points at 65,536 tokens and from $+2.67$ to $+1.97$ at 524,288 tokens. This calibration did not improve the positive-effect rule and does not determine the count in the main comparison (Appendix C.1). In a separate new-training-draw confirmation, Direct CTV admits 2 of 8 families for $+0.81$ points per decision, with 0 harmful admissions. The frozen shrinkage, dose-gating, and database-bound rules all retain the incumbent, so increased caution removes the observed gain (Appendix C.1).

6.2 How Scores Track the Measured Effect

Theorem 1 characterizes when a single dataset value transfers across states. Table 2 checks how the scores track measured effects at the recorded states. Always retaining the incumbent matches the effect sign on 102 of 144 views; no score-based policy exceeds that count, while measurement matches on 121. Across the 24 configuration–direction pairs, the two LESS variants have the highest rank correlation with the measured effect, but four of the seven score families order candidates against it. A ranking can distinguish candidates without supplying an effective admission threshold, and the shared-verdict bound identifies states where one dataset value cannot serve both.

Table 2: Score agreement with measured effects: action-sign matches over 144 decisions, and gap-sign matches and rank correlation over 24 configuration–direction pairs.

Score	Action matches sign	Gap sign matches	Spearman ρ
Teacher accuracy	67	11	-0.27
Trace judge	71	11	-0.21
Student NLL	78	13	-0.36
RSR	49	7	+0.40
Forward learnability	102	19	-0.36
LESS validation gradient	92	17	+0.63
LESS sign gradient	84	15	+0.63

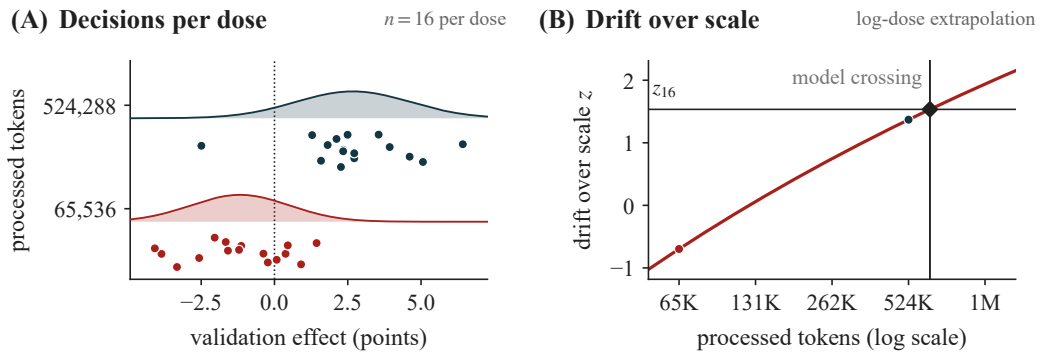

Figure 5: Two-dose transition model. (A) Effects for sixteen decisions at each dose and fitted normal curves. (B) Measured drift-to-scale ratios and extrapolation. The $z_{16} = \Phi^{-1}(1 - 1/16) \approx 1.53$ crossing means one expected harmful decision among sixteen; it lies beyond measured doses and has no uncertainty interval.

6.3 Where Choice Stops Paying

At 524,288 tokens fifteen of the sixteen effects are positive, the measurement admits all sixteen candidates, and every policy ties at that count. Figure 3 gives every family's effect at both receiving states and both doses. Writing $\mu(D)$ and $w(D)$ for the mean and standard deviation of the realized effects at dose D, and $z = \mu/w$ for their ratio, the behaviour of the comparison follows from one statement about those two numbers.

Proposition 2 (Dose transition)**.** *Let the N realized effects at dose D be independent draws from $\mathcal{N}(\mu, w^2)$, with the two states of a dataset drawn independently. The expected number of harmful decisions is $N\Phi(-z)$, a dataset flips sign with probability $2\Phi(z)\Phi(-z)$, and its expected sign-flip price under Corollary 1 is at most $2w\Phi(|z|)\psi(|z|)$, where $\psi(a) = \phi(a) - a\Phi(-a)$. No admission rule gains more than $w\psi(|z|) \le w\phi(z)/(1 + z^2)$ per decision over the better of admitting all and retaining all.*

The gain available to any choice of data therefore collapses faster than $\mathrm{e}^{-z^2/2}$ once the drift exceeds the scale. We call the dose where the expected number of harmful decisions falls to one, $z(D^\star) = z_N = \Phi^{-1}(1 - 1/N) = 1.53$ for $N = 16$, the transition dose. The measured ratio rises from -0.70 at 65,536 tokens to $+1.37$ at 524,288 tokens, with drift -1.18 and $+2.67$ points against scale 1.69 and 1.95. Figure 5 carries that movement, and the fitted model places the transition above 524,288 processed tokens, which is the largest dose we ran. Extending μ and w linearly in log dose under assumption (A3) puts the crossing at 637,000 tokens, and that figure comes from an extrapolation. Appendix C.6 gives the derivation of $D^\star$ and the in-sample consistency check against the same sixteen measured effects.

The cross-state study also carries the state effect that motivates the measurement. Reading the sixteen decisions at both receiving models gives effects that change sign between them, and a policy that scores the data alone returns one verdict for both readers. Appendix B.3 reports the cross-state study,

its registered gates, and the transfer regret a pipeline pays when it measures one model and deploys another.

6.4 The Price of the Measurement

The measurement trains continuations that the scores it replaces never train, so its price has to be stated in the same units it buys. A full measurement at the frozen benchmark costs 96 student continuations for twelve datasets. The data-level scores cost nothing beyond the scoring pass, and the one-step micro-model assay costs 24 continuations in total. Figure 9 prices the trade at intermediate budgets: with one repeated run the measurement realizes a smaller utility than the pooled correction that borrows from the other candidates, with two runs it passes it, and with four runs it reaches $+0.17$ points per decision. The measurement therefore has a range of useful budgets, and the corrections in Appendix C.5 are what make a short schedule competitive.

Two limits bound the claims. The evidence covers one task family and low-rank updates of 4B and 8B receiving models, up to 524,288 processed tokens, and the transition dose comes from an extrapolation. A third dose above the measured range and a calibrated stopping rule inside a running loop are the next tests, and the record of what failed is in Appendix D.

7 Conclusion

CTV measures a candidate dataset as an effect at its receiving state. The shared-verdict bound explains what a score can miss, and the paired continuation makes that effect observable at a counted cost. On the frozen benchmark, measurement improves the compared rankings and has positive mean utility over twelve configurations; database-level generalization remains unresolved. The follow-ups sharpen the distinction between ranking and admission: CTV family selection outperforms the LESS implementations but not uniform selection, while conservative gates forgo Direct CTV's gain on a new training draw. Choosing a better-ranked dataset and choosing a beneficial update remain separate judgments. The evidence covers text-to-SQL and low-rank updates; validating the extrapolated dose transition and admission inside a running loop remain open.

Reproducibility Statement

Section 6 defines the benchmarks, models, and metrics. Appendix B specifies data splits, optimization, evaluation, and comparison costs. Appendix A gives the mathematical derivations, and Appendix C reports the supporting results. Code and machine-readable analysis outputs will be released with the paper.

AI Use Statement

Generative artificial-intelligence (AI) tools assisted literature discovery, method and experiment design feedback, implementation, mathematical proof development, data processing, result analysis, figure design, and manuscript drafting and editing. Teacher language models generated the synthetic reasoning traces used in the training experiments; the experimental protocols specify their construction and evaluation.

References

Sina Alemohammad, Li Chen, Richard G. Baraniuk, and Zhangyang Wang. Not all synthetic data is yours to learn from, 2026. URL https://arxiv.org/abs/2605.31126.

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

Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q. Tran, and Mehran Kazemi. Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2408.16737.

BIRD Team. LiveSQLBench: A dynamic and contamination-free benchmark for evaluating LLMs on real-world text-to-SQL tasks. https://github.com/bird-bench/livesqlbench, 2024.

Yuxi Chen, Hamza Golubovic, Han Tong, Arian Maleki, and Andrew Ilyas. Data attribution via sketched metadifferentiation, 2026a. URL https://arxiv.org/abs/2609.15044.

Zhuofan Chen, Ziqian Jiao, Yikai Cui, Zhixin Cai, Jun Bai, and Wenge Rong. CircuitLens: Reasoning circuits as data selection signals for reinforcement learning with verifiable rewards, 2026b. URL https://arxiv.org/abs/2609.07183.

Rishit Dagli, Abir Harrasse, Luke Zhang, Florent Draye, Amirali Abdullah, Bernhard Schölkopf, and Zhijing Jin. STRIDE: Training data attribution via sparse recovery from subset perturbations, 2026. URL https://arxiv.org/abs/2606.05165.

Cuong Dang, Hoang Anh Just, and Ruoxi Jia. Capacity-dependent effects of data selection for reasoning. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2608.13721.

Junwei Deng, Pingbang Hu, Suliang Jin, Hao Lu, Jiachen T. Wang, Shichang Zhang, and Jiaqi W. Ma. How faithful is Trajectory-Based data attribution? error sources, remedies, and practical guidelines, 2026. URL https://arxiv.org/abs/2605.18814.

Meng Ding, Zeqing Zhang, Di Wang, and Lijie Hu. In-run data Shapley for Adam optimizer, 2026. URL https://arxiv.org/abs/2602.00329.

Qiang Gao, Zhenping Li, Anqi Zhuo, Yingxiao Zhao, Weibo Geng, and Xiaosong Li. SemanticAgent: A Semantics-Aware framework for Text-to-SQL data synthesis, 2026. URL https://arxiv.org/abs/2604.21414.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Jinhui Guo. Delayed optimizer-state transport shapes short-horizon training decisions, 2026. URL https://arxiv.org/abs/2608.24593.

Keisuke Hirano and Jack R. Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701, 2009. doi: 10.3982/ECTA6630. URL https://doi.org/10.3982/ECTA6630.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Michael Y. Hu, Apurva Gandhi, Kyunghyun Cho, Tal Linzen, and Pratyusha Sharma. Always learning, always mixing: Efficient and simple data mixing all the time, 2026a. URL https://arxiv.org/abs/2605.15220.

Zhengyu Hu, Zheyuan Xiao, Linxin Song, Fengqing Jiang, Yuetai Li, Zhihan Xiong, Yue Liu, Junhao Lin, Yao Su, Lijie Hu, Kaize Ding, Teng Xiao, and Radha Poovendran. The strongest teacher is not always the best teacher: Student-centric answer selection, 2026b. URL https://arxiv.org/abs/2605.26872.

Jiahao Huang, Fei Cheng, Junfeng Jiang, and Akiko Aizawa. Tailoring the curriculum: Student-centered reasoning distillation via dynamic data-model compatibility, 2026. URL https://arxiv.org/abs/2605.29229.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9525–9587. PMLR, 2022. URL https://proceedings.mlr.press/v162/ilyas22a.html.

Pratik Kakkar, Chandra Dhir, Ravi Shankar, Pareekshit Reddy Gaddam, and Anup Shirgaonkar. ReToolSQL: Agentic reinforcement learning for robust Text-to-SQL, 2026. URL https://arxiv.org/abs/2608.27796.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018. doi: 10.3982/ECTA13288. URL https://doi.org/10.3982/ECTA13288.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Jin Hwa Lee, Matthew Smith, Maxwell Adam, and Jesse Hoogland. Influence dynamics and stagewise data attribution, 2025. URL https://arxiv.org/abs/2510.12071.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. Can LLM already serve as a database interface? A BIg Bench for large-scale database grounded text-to-SQLs. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/83fc8fab1710363050bbd1d4b8cc0021-Abstract-Datasets_and_Benchmarks.html.

Qihao Lin, Guanxu Chen, Dongrui Liu, and Jing Shao. PRISM: Preference-Aware influence function based data selection method for efficient Fine-Tuning, 2026. URL https://arxiv.org/abs/2605.21422.

Kang Liu and Suyan Li. Finite-horizon input-output dynamics of minibatch perturbations in AdamW, 2026. URL https://arxiv.org/abs/2608.19762.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. TSDS: Data selection for task-specific model finetuning, 2024. URL https://arxiv.org/abs/2410.11303.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Xiaonan Luo, Yue Huang, Kehan Guo, Ping He, Chuan Zou, Ting Hua, and Xiangliang Zhang. Learning from synthetic data without model collapse in iterative instruction tuning, 2026a. URL https://arxiv.org/abs/2607.17043.

Yuzhang Luo, Chenpeng Wang, Jianhui Chen, and Liangming Pan. From reweighting to rewriting: Unlocking the intervention effects of influential samples in training data attribution, 2026b. URL https://arxiv.org/abs/2609.02771.

Haoyuan Ma, Hengwei Liu, Linjuan Wu, Yongliang Shen, and Weiming Lu. LIMIT: Less is more for instruction tuning in Text-to-SQL, 2026. URL https://arxiv.org/abs/2609.24186.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72 (4):1221–1246, 2004. doi: 10.1111/j.1468-0262.2004.00530.x. URL https://doi.org/10.1111/j.1468-0262.2004.00530.x.

Charles F. Manski and Aleksey Tetenov. Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *Journal of Statistical Planning and Inference*, 137(6):1998–2010, 2007. doi: 10.1016/j.jspi.2006.06.037. URL https://doi.org/10.1016/j.jspi.2006.06.037.

Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt, 2022. URL https://arxiv.org/abs/2206.07137.

José Luis Montiel Olea, Chen Qiu, and Jörg Stoye. Decision theory for treatment choice problems with partial identification. *The Review of Economic Studies*, 2026. doi: 10.1093/restud/rdag015. URL https://doi.org/10.1093/restud/rdag015.

Yuto Nishida, Hirokazu Kiyomaru, Yusuke Oda, Takashi Kodama, Chaoran Liu, Daisuke Kawahara, Yusuke Miyao, Max Müller-Eberstein, and Masaru Isonuma. Measuring Task-Agnostic training data influence across language model pretraining. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2608.13515.

Adrian Nyakairu and Hongfu Liu. From detrimental to beneficial: Dynamic influence-based valuation and editing, 2026. URL https://arxiv.org/abs/2608.22522.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Haolong Qian, Xianliang Yang, Yinuo ma, Lirong Che, Feng Lu, Ye Guo, Lei Song, Jiang Bian, and Chun Yuan. The Quality-Utility paradox: Why High-Reward data impairs small model mathematical reasoning, 2026. URL https://arxiv.org/abs/2606.16152.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350. URL https://doi.org/10.1037/h0037350.

Yinan Shen, Ziao Yang, and Hongfu Liu. Validation-Induced shapley shifts: How validation structure distorts data valuation, 2026. URL https://arxiv.org/abs/2607.03675.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1): 70–81, 2009. doi: 10.1016/j.jeconom.2009.02.013. URL https://doi.org/10.1016/j.jeconom.2009.02.013.

Peng Sun, Yi Yang, Antong Zhang, Chunxiao Li, Yanbo Wang, Dianbo Liu, Xin Chen, Kai Yu, Lu Chen, and Tianfan Fu. Data-DPO: Direct preference optimization for target model data selection in LLM post-training, 2026. URL https://arxiv.org/abs/2608.16926.

Noel Thomas. No subspace to track: Non-identifiability and optimizer state in low-rank training, 2026. URL https://arxiv.org/abs/2607.05872.

Xiao Tian, Jue Fan, Rachael Hwee Ling Sim, Zixuan Wang, Nancy F. Chen, and Bryan Kian Hsiang Low. Is Data Shapley not better than random in data selection? ask NASH. In *Forty-third International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.10684.

Bing Wang, Rui Miao, Chen Shen, Shaotian Yan, Kaiyuan Liu, Ximing Li, Xiaosong Yuan, Sinan Fan, Jun Zhang, and Jieping Ye. On the step length confounding in LLM reasoning data selection, 2026a. URL https://arxiv.org/abs/2604.06834.

Fangxin Wang, Peyman Baghershahi, Langzhou He, Henry Peng Zou, Sourav Medya, and Philip S. Yu. Filter-then-weight: Online data selection and reweighting for LLM fine-tuning, 2026b. URL https://arxiv.org/abs/2604.00001.

Jiachen T. Wang, Dawn Song, James Zou, Prateek Mittal, and Ruoxi Jia. Capturing the temporal dependence of training data influence, 2024. URL https://arxiv.org/abs/2412.09538.

Shaobo Wang, Xuan Ouyang, Tianyi Xu, Yuzheng Hu, Jialin Liu, Guo Chen, Tianyu Zhang, Junhao Zheng, Kexin Yang, Xingzhang Ren, Dayiheng Liu, and Linfeng Zhang. OPUS: Towards efficient and principled data selection in large language model pre-training in every iteration. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026c. URL https://icml.cc/virtual/2026/poster/61926.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Jing Xiao, Xinhai Chen, Qinglin Wang, Menghan Jia, Zhiquan Lai, Dongsheng Li, Jie Liu, and Tiejun Li. Gradient-update mismatch: Rethinking conflict-free training of physics-informed neural networks, 2026. URL https://arxiv.org/abs/2609.01558.

Qinyang Xu. Stored in optimizer state, valued by later training: A causal account of subliminal trait transfer, 2026. URL https://arxiv.org/abs/2608.20442.

Yuxin Yang, Aoxiong Zeng, and Xiangquan Yang. The Long-Term effects of data selection in LLM Fine-Tuning, 2026. URL https://arxiv.org/abs/2605.30537.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://arxiv.org/abs/2207.01206.

Tianrun Yu, Kaixiang Zhao, Chih-Chun Chen, Amanda Hughes, Taylor W. Killian, Fenglong Ma, Weitong Zhang, and Porter Jenkins. LARK: Learnability-grounded trajectory selection for efficient reasoning distillation, 2026. URL https://arxiv.org/abs/2605.30651.

Zichun Yu, Spandan Das, and Chenyan Xiong. MATES: Model-aware data selection for efficient pretraining with data influence models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/c4bec0d2fd217e6c2c3eafeced432582-Abstract-Conference.html.

Siqi Zeng, Christopher Jung, Rui Li, Zhe Kang, Ming Li, Nima Noorshams, Zhigang Wang, Fuchun Peng, Han Zhao, and Xue Feng. Convex dataset valuation for Post-Training. In *Forty-third International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.16704.

Dewu Zheng, Ruizhe Ye, Yanlin Wang, Yang Ye, Hongyu Zhang, Ensheng Shi, Xilin Liu, Yuchi Ma, Jianxing Yu, and Zibin Zheng. SWE-Prime: Fewer trajectories, better performance, 2026. URL https://arxiv.org/abs/2608.27449.

Ruiqi Zhong, Tao Yu, and Dan Klein. Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 396–411, 2020. doi: 10.18653/v1/2020.emnlp-main.29.

Yuxuan Zhu, Tengjun Jin, Yoojin Choi, and Daniel Kang. Human-level text-to-SQL via reinforcement learning on verified data, without pipeline engineering, 2026. URL https://arxiv.org/abs/2603.20004.

A PROOFS

A.1 PROOF OF THEOREM 1

Theorem 1 (Transferable dataset values). *Let $F : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ be the response table on finite sets of states and datasets. Fix anchors $x_0 \in \mathcal{X}$ and $a_0 \in \mathcal{A}$, and define*

$$\Delta(x, a) = F(x, a) - F(x, a_0) - F(x_0, a) + F(x_0, a_0).$$

Every anchored interaction vanishes exactly when $F(x, a) = b(x) + g(a)$ for some functions b and g. The components $b(x) = F(x, a_0)$ and $g(a) = F(x_0, a) - F(x_0, a_0)$ are then the unique pair with $g(a_0) = 0$. When every $|\Delta(x, a)| \le \varepsilon$, those components reconstruct the table to within ε.

Proof. For the anchors (x_0, a_0) in the theorem, set $b(x) = F(x, a_0)$ and $g(a) = F(x_0, a) - F(x_0, a_0)$. Rearranging the definition of the anchored interaction gives $F(x, a) = b(x) + g(a) + \Delta(x, a)$. Thus zero interaction is equivalent to the additive representation, and $|\Delta(x, a)| \le \varepsilon$ gives the stated uniform approximation. Conversely, an additive F has zero interaction by cancellation. If $g(a_0) = 0$, evaluation at a_0 forces $b(x) = F(x, a_0)$; evaluation at x_0 then forces the displayed $g(a)$, proving uniqueness. □

A.2 PROOF OF COROLLARY 1

Corollary 1 (Sign-flip price). *At state i, admitting realizes τ_i and retaining realizes zero. A state-free rule reads identically distributed inputs at the two states and admits with the same probability $p \in [0, 1]$. Suppose $\tau_1 > 0 > \tau_2$. Every state-free rule has summed expected regret $(1 - p)|\tau_1| + p|\tau_2| \ge \min(|\tau_1|, |\tau_2|)$ over the two states, and the case $\tau_1 < 0 < \tau_2$ exchanges the two effects. A deterministic state-free rule has regret at least $\min(|\tau_1|, |\tau_2|)$ at one of the states, and every randomized one has worst-state expected regret $\max\{(1 - p)|\tau_1|, p|\tau_2|\} \ge |\tau_1||\tau_2|/(|\tau_1| + |\tau_2|)$, with equality at $p = |\tau_1|/(|\tau_1| + |\tau_2|)$.*

Proof. Identically distributed inputs induce the same admission probability p at both states. Let $a = \tau_1 > 0$ and $b = -\tau_2 > 0$. Expected regrets are $(1 - p)a$ and pb. Their sum is affine in p and has minimum $\min(a, b)$ at an endpoint. For deterministic p, one state incurs regret a or b. The worst-state regret for randomized p is minimized when $(1 - p)a = pb$, namely at $p = a/(a + b)$, and equals $ab/(a + b)$. Exchanging the two states covers the opposite sign order. □

A.3 PROOF OF PROPOSITION 1

Proposition 1 (Seed complexity). *For a fixed effect τ, let $d_j = \tau + e_j$, where $e_1, \ldots, e_n$ are independent, mean-zero, sub-Gaussian errors with variance proxy σ^2, and let $\bar{d} = n^{-1} \sum_{j=1}^{n} d_j$. The rule that admits when $\bar{d} > 0$ errs in sign with probability at most $\exp(-n\tau^2/2\sigma^2)$. Hence, for a floor $\tau_{\min} > 0$, $n \ge 2\sigma^2 \log(1/\delta)/\tau_{\min}^2$ runs decide the sign of every effect with $|\tau| \ge \tau_{\min}$ with error at most δ, and the expected regret is at most $|\tau| \exp(-n\tau^2/2\sigma^2) \le \sigma/\sqrt{\mathrm{e}\, n}$ for every τ.*

Proof. The independent sub-Gaussian errors imply that $\bar{d} - \tau$ is sub-Gaussian with variance proxy σ^2/n. For $\tau > 0$, a wrong sign requires $\bar{d} - \tau \le -\tau$; for $\tau < 0$, it requires the symmetric upper-tail event. Either tail is at most $\exp[-n\tau^2/(2\sigma^2)]$. Substituting $|\tau| \ge \tau_{\min}$ and solving for n proves the seed requirement. An erroneous action loses $|\tau|$, so expected regret is at most $|\tau| \exp[-n\tau^2/(2\sigma^2)]$. Its maximum over $|\tau| \ge 0$ occurs at $|\tau| = \sigma/\sqrt{n}$ and is $\sigma/\sqrt{en}$. □

A.4 PROOF OF PROPOSITION 2

Proposition 2 (Dose transition). *Let $z = \mu/w$ for $w > 0$, and let ϕ and Φ denote the standard normal density and cumulative distribution function. Let the N realized effects at dose D be independent draws from $\mathcal{N}(\mu, w^2)$, with the two states of a dataset drawn independently. The expected number of harmful decisions is $N\Phi(-z)$, a dataset flips sign with probability $2\Phi(z)\Phi(-z)$, and its expected sign-flip price under Corollary 1 is at most $2w\Phi(|z|)\psi(|z|)$, where $\psi(a) = \phi(a) - a\Phi(-a)$. No admission rule gains more than $w\psi(|z|) \le w\phi(z)/(1 + z^2)$ per decision over the better of admitting all and retaining all.*

Proof. Linearity of expectation gives $N\Phi(-z)$ harmful decisions. Independence of the two receiving-state effects gives the sign-flip probability $2\Phi(z)\Phi(-z)$. Normal integration gives $\mathbb{E}[\tau^-] = w\psi(z)$ and $\mathbb{E}[\tau^+] = w\psi(-z)$, where $\tau^- = \max(-\tau, 0)$ and $\tau^+ = \max(\tau, 0)$. On a sign flip the smaller absolute effect is at most the magnitude of the effect whose sign opposes μ. Independence therefore bounds the expected sign-flip price in Corollary 1 by $2w\Phi(|z|)\psi(|z|)$. An oracle that admits exactly positive effects gains at most $w\psi(|z|)$ per decision over the better constant action. The normal-tail bound $\Phi(-a) \geq a\phi(a)/(1 + a^2)$ for $a \geq 0$ further bounds this gain by $w\phi(z)/(1 + z^2)$. □

B Experimental Protocol

The experiments measure the value of a fixed token budget of training data at a declared student checkpoint. Teacher datasets, trained students, optimizer seeds, evaluation queries, and repeated responses retain their distinct statistical roles.

B.1 Result hierarchy and statistical units

The studies measure BIRD text-to-SQL execution accuracy: a generated query succeeds when it returns the gold result under the declared SQLite evaluator (Li et al., 2023; Zhong et al., 2020). Training traces come from BIRD training databases disjoint from all evaluation databases. The custom evaluation split has 161 development, 179 selection, and 441 validation questions from 20 databases. Selection questions determine admissions; validation questions measure the resulting utility. Candidate and incumbent continuations share the saved receiving state, updater seed, question, and generation address. Differences are paired before aggregation. An evaluation question is a repeated readout of a trained student, not a separate training replicate.

The cross-state study has eight candidate families—six trace producers and two retrieval arms—at each of two receiving models and two token doses. The incumbent is the merged retrieval pool from the same draw. Three updater seeds underlie each candidate–incumbent effect. Its registered uncertainty procedure resamples producer families jointly across receiving models, seeds, and policies, with validation questions clustered by database. It reports one-sided lower bounds at level 0.05. The frozen decision benchmark has twelve training configurations, each observed under two cross-fit question-role directions and six evaluation criteria, yielding 144 views of twelve trained configurations rather than 144 independent fits. Four optimizer seeds replicate each configuration. Its 384 evaluation roots form 192 matched pairs within eleven databases; the action and adjudication roles are exchanged and averaged. The six criteria are the three deployment endpoints separately and their uniform mixture, worst endpoint, and total-variation-robust objective at radius 0.25.

The motivation and state-intervention readouts in Figure 2 have separate populations. Four natural BIRD trace sets each have 48 source cases, two generation attempts per case, and 96 packed addresses. Qwen3-1.7B receives 32,768 processed tokens per retrieval condition; seeds 2719 and 3253 produce sixteen fine-tuned models, whose case-level responses are paired across three evaluation settings. The AdamW order/reset intervention fixes panels B and D before mechanism training, selecting the trace set with the largest positive source-success gain among those with a negative minimal-setting response. It uses 192 selection and 192 validation cases per model; two later seeds, 4517 and 6421, bring it to 32 models. The Qwen3.5-2B method-held-out comparison uses panel D and seed 6421. Its state intervention crosses those panels with UU, US, SU, SS block orders, comparing sixteen continuous-AdamW runs with sixteen second-moment-rebase runs. These cases are repeated observations of the trained models. The mechanism fits use rank-8 LoRA on the four attention projections, alpha 16, zero dropout, and AdamW at 2×10^{-4} with the epsilon, decay and clipping settings below. Each block processes 16,384 tokens in eight steps. Natural-response generation uses temperature 0.7, top-p 0.9, a 160-token cap and seed 64013. Natural-study intervals resample trace sets before paired cases and seeds; switchback intervals preserve panel–seed cells and resample paired cases.

B.2 Receiving states and training

The frozen benchmark uses Qwen3-4B at revision `1cfa9a72` and Qwen3-8B at revision `b968826d`. Model, tokenizer, evaluator, and dependency revisions are fixed within each paired block. A receiving state names the saved LoRA factors, AdamW moments, step, and random state in the factorization

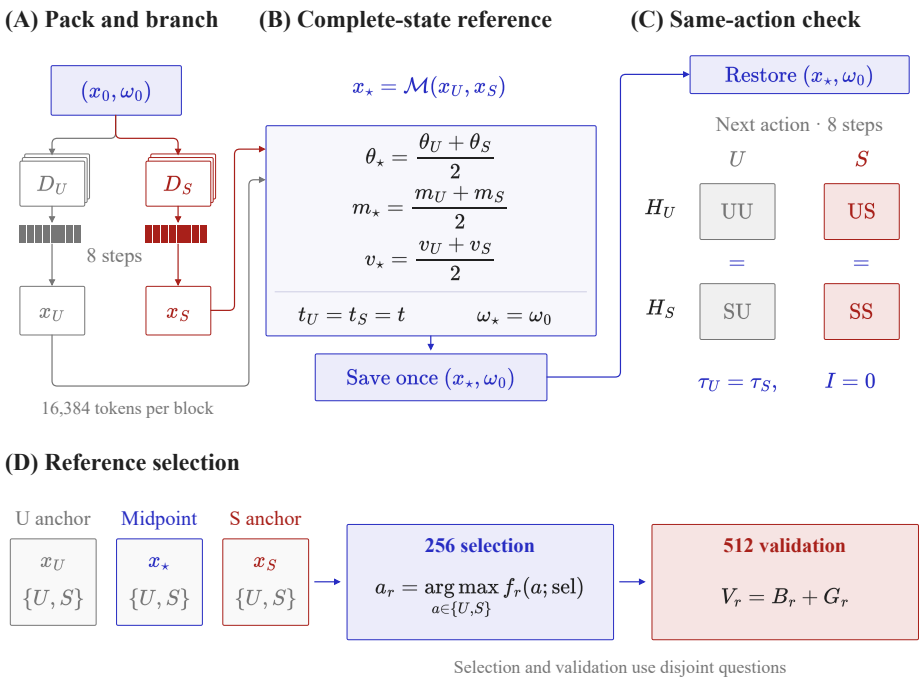

Figure 6: Complete-state restoration and reference selection. Paired continuations share parameters, optimizer moments, clock, and randomness; selection and validation questions are disjoint.

Table 3: Studies and supporting evidence.

Study	Question and location
Cross-state study	Whether measurement ranks better and whether harm depends on the state (Figures 4 and 7)
Menu replication	Whether measured choice raises official accuracy (App. D.2)
Decision benchmark	Which admission policy helps the next update (Table 6, App. C.2)
Debiased measurement	What a smaller seed budget recovers (Figure 9, App. C.5)
Matched-budget grid	Whether a score picks better inside one pool (App. D.2)
Selector benchmark	Whether published selectors beat their controls (App. D.2, App. D.2)
Natural trace sets	How teacher and student effects relate (App. B.1)
State interventions	What complete-state restoration controls (App. C.7, App. C.8)
Reference comparison	Which receiving state to measure from (App. C.9)
Score transfer	How far each score sits from the measured effect (App. C.3)
Evaluation scope	Where the evaluator bounds the claims (App. D)

actually written. The paired continuations restore those bytes. The effect therefore belongs to that saved low-rank state; an alternative factorization of the same represented weight update is outside this identity claim (Thomas, 2026).

In the cross-state study, each state starts from its base weights and fresh AdamW moments. A continuation updates rank-16 LoRA adapters with alpha 32 on the query, key, value, output, gate, up, and down projections at learning rate 2×10^{-4} and AdamW $\beta = (0.9, 0.95)$. Each registered dose takes 32 optimizer steps, processing either 65,536 or 524,288 tokens; Qwen3-8B uses accumulation and checkpointing without changing the effective batch or token count. The incumbent continuation is shared among the eight candidate comparisons for a given state, dose, and seed. Thus the two original doses require 108 unique continuations: $2 \times 2 \times 3 \times (8 + 1)$.

The frozen benchmark trains the candidate and incumbent under each configuration at seeds 2719, 3253, 4517, and 6421, giving 96 unique students. All fits adapt the four attention projections with LoRA, rank 8 or 32 and alpha twice the rank. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$,

Table 4: Twelve training configurations, each repeated at four seeds for candidate and incumbent.

Trace family	Scale	Rank	Alpha	Learning rate	Steps	Tokens
Bounded verifier	4B	32	64	1×10^{-4}	4	8,192
Bounded verifier	4B	8	16	2×10^{-4}	16	32,768
Bounded verifier	8B	32	64	1×10^{-4}	32	65,536
Bounded verifier	8B	32	64	1×10^{-4}	4	8,192
Runtime state	4B	32	64	1×10^{-4}	16	32,768
Runtime state	4B	32	64	1×10^{-4}	2	4,096
Runtime state	4B	8	16	2×10^{-4}	4	8,192
Runtime state	8B	8	16	2×10^{-4}	32	65,536
Solution verifier	4B	8	16	2×10^{-4}	2	4,096
Solution verifier	8B	32	64	1×10^{-4}	16	32,768
Solution verifier	8B	8	16	2×10^{-4}	2	4,096
Solution verifier	8B	8	16	2×10^{-4}	32	65,536

zero weight decay, and gradient clipping at 1. Learning rate is $2 \times 10^{-4}\sqrt{8/r}$ at rank r; each step processes 2,048 tokens. The base checkpoint is restored between fits, and supervision scores the current assistant turn only. Packing retains complete traces, failures, and retries, orders fixed addresses with a prespecified salt, and fills the token budget before target effects are read.

B.3 Registered decisions and readouts

The cross-state registration preceded producer generation, training, and readout. One pre-readout amendment replaced its 40-dataset design with the sixteen first-draw candidate–state decisions at each dose. A dataset score admits its top half of the pool, while Direct CTV admits when the seed-mean selection effect is positive. Scores and actions were fixed before validation readout. The registration's four gates use all 781 custom-split questions: between-dataset effect standard deviation above 1.5 points (G1); at least three datasets passing both execution and judge filters while losing two or more points (G2); at least two cross-state sign-changing families (G3); and absolute effect–score Spearman correlation below 0.5 for both LESS and Data-DPO (G4). At 65,536 tokens, the measured quantities are 0.00 points (G1, fails), 3 harmful filter-passing datasets (G2, passes), 2 sign-changing families (G3, passes), and correlations $+0.24$ and -0.24 (G4, passes). At 524,288 tokens they are 1.33 (fails), 0 (fails), 1 (fails), and -0.17 and -0.58 (fails). The completed readout covers all 108 continuations; the 1,048,576-token extension has no complete paired decision.

The two-dose comparison retains the registered criteria above. Its score comparators use the evaluated dataset-level adaptations, including receiving-state low-rank gradients for TracIn and TSDS. Incomplete scores have no result. An exploratory seven-seed extension remeasures the binary verifier, persistent repair, and structure-relevant retrieval at both receiving states. Opposite mean signs persist for the first two families, but their seed-bootstrap intervals overlap zero at both states; the extension does not establish stable reversals.

Table 5 reports completed policy comparisons, whose admission counts can differ. Accuracy is the fraction of actions matching measured effect signs; utility is validation accuracy relative to the incumbent, in percentage points. Gap is Direct CTV minus the row, and Bound is its one-sided 0.05 lower bound from the registered family bootstrap with validation questions clustered by database.

Filters and admit-all must show a gap of at least 1.0 point with a positive bound; the LESS-style proxy and Data-DPO require at least 0.5 point with a bound above -0.25. Teacher accuracy was scored after the first readout and is exploratory. Dashes mark self-comparisons or policies without a registered criterion.

B.4 Official benchmark and cost

Official BIRD development accuracy is measured separately on all 1,534 questions with a frozen schema-only prompt, no example rows, 256 generated tokens, temperature zero, top-p one, disabled sampling, and the official SQLite execution-accuracy evaluator. The untrained Qwen3.5-2B and

Table 5: Cross-state decisions at two token doses; dashes mark inapplicable entries.

	65,536 tokens					524,288 tokens				
Policy	Acc.	Utility	Gap	Bound	Criterion	Acc.	Utility	Gap	Bound	Criterion
Fixed actions										
Admit every dataset	0.31	-1.18	+0.68	+0.05	not met	0.94	+2.67	+0.00	+0.00	not met
Retain the incumbent	0.69	+0.00	-0.50	-1.69	–	0.06	+0.00	+2.67	+0.09	–
Filters										
Teacher accuracy	0.56	-0.57	+0.07	-0.68	expl.	0.56	+1.35	+1.32	-0.09	expl.
Execution correctness	0.56	-0.91	+0.41	-0.73	not met	0.44	+0.92	+1.76	+0.04	met
Trace judge	0.44	-0.92	+0.42	-0.47	not met	0.44	+1.02	+1.66	+0.10	met
Gradient and probing scores										
LESS-style proxy	0.31	-0.91	+0.41	-0.22	not met	0.44	+1.27	+1.41	+0.09	met
Data-DPO	0.31	-0.82	+0.32	-0.34	not met	0.44	+1.18	+1.50	+0.21	met
Measurement at the state										
Debiased measurement	0.31	-1.18	+0.68	+0.05	–	0.94	+2.67	+0.00	+0.00	–
Direct CTV	0.38	-0.50	–	–	–	0.94	+2.67	–	–	–

Table 6: Admission policies on 144 frozen decisions under each policy's own rule.

Policy	Admit rate (%)	Harmful	$\Pr(>0)$	Utility (pp)	Amortized fits/view
Fixed actions					
Keep the incumbent	0.0	0	0.000	+0.00	0.00
Take every candidate	100.0	101	0.001	-1.66	0.00
Random selection	29.2	29	0.001	-0.48	0.00
Teacher outcome					
Teacher accuracy (Li et al., 2023)	66.7	65	0.024	-1.22	0.00
Dataset-level scores					
Trace judge (Zhu et al., 2026)	33.3	39	0.000	-1.24	0.00
Student NLL (Yu et al., 2026)	33.3	36	0.000	-0.43	0.00
RSR (Sun et al., 2026)	83.3	86	0.002	-1.59	0.00
†LESS validation gradient (Xia et al., 2024)	50.0	41	0.325	<u>-0.05</u>	0.00
LESS sign gradient (Xia et al., 2024)	58.3	51	0.092	-0.28	0.00
Measured on a student					
1.7B transport student (Yu et al., 2024; Mindermann et al., 2022)	37.5	37	0.001	-0.52	0.00
Micro-model CTV (Yu et al., 2026)	34.0	<u>34</u>	0.058	-0.38	0.17
Direct CTV	29.2	11	0.99995	**+0.17**	0.67

Qwen3.5-9B scale anchors score 427/1,534 (27.84%) and 784/1,534 (51.11%), respectively; neither is a receiving model in the two main studies. Custom-split results and official development accuracy have different evaluation populations.

The frozen benchmark charges Direct CTV for its 96 receiving-scale continuations. The three endpoints are candidate runtime, incumbent runtime, and the minimal runtime; each is an evaluation setting for the same trained student. Every student answers 384 held-out roots at these endpoints, or 1,152 generations per fit; held-out generation contributes 50% of its accounted cost. The six dataset-level scores share one candidate-trace scoring pass, costed at 0.003 relative to Direct CTV. The 1.7B transport atlas costs 0.16, the 24-fit one-seed micro-model assay 0.24, and calibrated supported transport 0.65. Candidate fits that the pipeline admits can also serve as deployed updates, so the decision and downstream fit costs overlap.

C Additional Results on the Main Decisions

In Table 6, utility is the mean validation gain per decision; $\Pr(>0)$ is the fraction of configuration-bootstrap draws with positive mean utility. Fits/view divides receiving-scale continuations by 144 decision views. Direct CTV has 30 beneficial admissions, 11 harmful admissions, and one tie. Random selection averages all subsets of that size in every resample. The dagger identifies the strongest cited rival; underlining marks the second-best entry. Comparator adaptations are described in Appendix D.1.

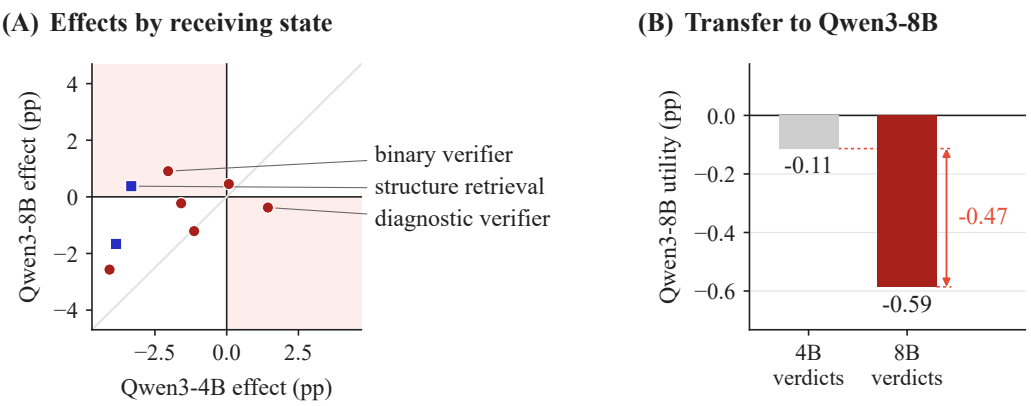

Figure 7: Cross-state transfer at 65,536 tokens. (A) Candidate-minus-incumbent effects at Qwen3-4B and Qwen3-8B; shading marks sign reversals. (B) Qwen3-8B utility from admissions chosen at either state.

The low-dose sign changes in Figure 7 describe effects at two receiving states, with seed noise larger than the between-dataset variance at that dose. Using Qwen3-4B measurements for Qwen3-8B admissions yields -0.11 points, compared with -0.59 from Qwen3-8B measurements. The registered lower bounds distinguish observed gaps from supported ones; the low-dose Direct CTV mean itself is -0.50 points relative to retaining the incumbent.

C.1 Admission thresholds

The primary rule admits a candidate when its selection-set mean effect is positive. Threshold calibration on the 161 development questions was proposed after inspecting validation outcomes and has no fresh confirmation. It reduces Direct CTV validation utility from -0.50 to -0.73 points at 65,536 tokens and from $+2.67$ to $+1.97$ at 524,288 tokens. A separate post hoc paired-database lower-bound rule and a shrinkage-plus-dose gate retain the incumbent in every low-dose decision, yielding zero utility; at the larger dose the lower-bound rule again retains every incumbent and the gated rule admits every candidate. These checks supply no confirmed improvement to the positive-effect rule.

A later confirmation freezes the shrinkage and dose-gating rules before reading a new training draw, using eight freshly generated candidate families at Qwen3-4B, 65,536 tokens and three updater seeds. The 27 candidate and shared-incumbent continuations reuse the existing evaluation database roles, so this tests a new training draw rather than new evaluation domains. Direct CTV admits 2/8, with utility $+0.81$ points and 0 harmful admissions. Shrinkage, the dose gate, their combination and the paired-database lower-bound rule all abstain. The primary gated-shrinkage-minus-direct contrast is -0.81 points, with one-sided lower bound -2.61 from 10,000 paired draws over families, databases and updater seeds. It fails the prespecified improvement criterion of at least 0.5 points with a positive lower bound; all methods share the same continuations. The completed analysis selects a fixed eligible execution group and retains incomplete attempts separately.

C.2 Uncertainty across configurations and databases

Table 7 shows the distinction between repeated decision views, training configurations, and source databases. Each level uses 20,000 resampling draws with seed 20260922. Database draws resample whole matched question-pair clusters and recompute each policy's decisions. Policies that never switch have zero-width decision- and configuration-level intervals; supported transport can switch after database resampling. Direct CTV gains $+0.1736$ points per decision on the observed benchmark, or $+0.5952$ per admission. Its interval excludes zero when decisions or twelve configurations are resampled. With eleven databases as clusters, the interval is $[-0.15, +0.21]$ points and 0.736 of

Table 7: Utility uncertainty across 144 decisions, 12 training configurations, or 11 databases; intervals are 95% bounds in percentage points.

	Utility	Decision level		Configuration level		Database level	
Policy	(pp)	Interval	Pr(> 0)	Interval	Pr(> 0)	Interval	Pr(> 0)
Keep the incumbent	+0.000	[+0.00, +0.00]	0.000	[+0.00, +0.00]	0.000	[+0.00, +0.00]	0.000
Take every candidate	-1.658	[-2.30, -1.07]	0.000	[-3.23, -0.39]	0.001	[-2.24, -1.27]	0.000
Teacher accuracy	-1.225	[-1.84, -0.67]	0.000	[-2.85, -0.01]	0.024	[-1.61, -1.01]	0.000
Trace judge	-1.236	[-1.85, -0.69]	0.000	[-2.88, -0.05]	0.000	[-1.54, -0.98]	0.000
Student NLL	-0.434	[-0.68, -0.22]	0.000	[-1.06, -0.01]	0.000	[-0.77, -0.15]	0.001
RSR	-1.592	[-2.24, -1.01]	0.000	[-3.19, -0.28]	0.002	[-2.02, -1.35]	0.000
Forward learnability	+0.000	[+0.00, +0.00]	0.000	[+0.00, +0.00]	0.000	[+0.00, +0.00]	0.000
LESS validation gradient	-0.054	[-0.19, +0.08]	0.208	[-0.29, +0.16]	0.325	[-0.41, +0.21]	0.263
LESS sign gradient	-0.283	[-0.53, -0.06]	0.004	[-0.78, +0.09]	0.092	[-0.70, +0.03]	0.035
Supported transport	+0.000	[+0.00, +0.00]	0.000	[+0.00, +0.00]	0.000	[-0.02, +0.03]	0.424
1.7B transport student	-0.519	[-0.92, -0.21]	0.000	[-1.16, -0.08]	0.001	[-0.74, -0.33]	0.000
Micro-model CTV	-0.383	[-0.71, -0.12]	0.001	[-1.03, +0.04]	0.058	[-0.73, -0.00]	0.024
Direct CTV	+0.174	[+0.09, +0.27]	1.000	[+0.07, +0.28]	1.000	[-0.15, +0.21]	0.736

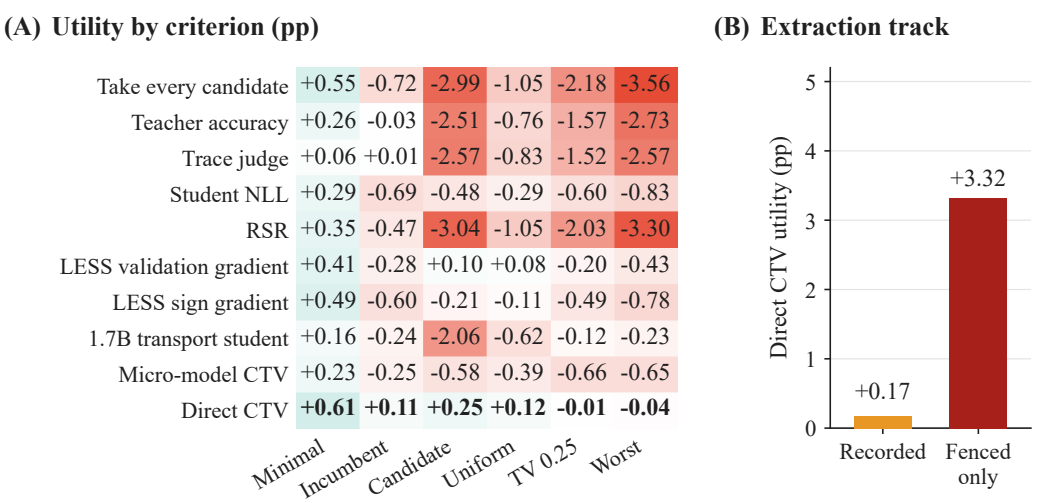

Figure 8: Frozen-decision utility. (A) Six evaluation criteria. (B) Direct CTV under recorded and fenced-only SQL extraction on the same responses.

draws are positive. The positive gain is thus unresolved at the coarsest available level; the benchmark records no producer-population label for another such resampling analysis.

C.3 Score agreement

Action agreement uses the 144 decision views. Gap signs and rank correlations average the six criteria within each configuration and cross-fit direction, giving 24 pairs.

C.4 Evaluator dependence

The recorded extractor reproduces all 160,142 stored executed queries on re-extraction, with 0 mismatches. 13,573 responses lack SQL code fences, and 2,165 of the 43,732 successes reach the executor through the fallback extraction rule. Counting those responses as failures gives Direct CTV $+3.32$ points, with 71 admissions and 6 harmful admissions; its recorded-track utility is $+0.17$ points. Both readings are positive, but the measured utility includes an answer-format effect and is conditional on the extraction rule.

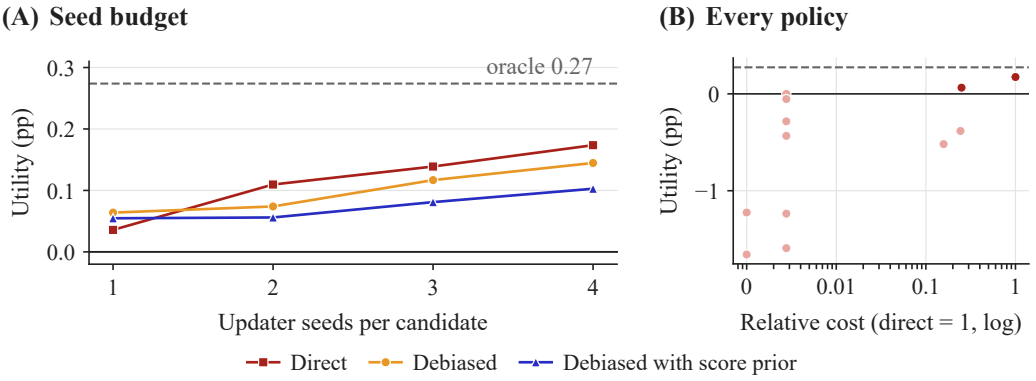

Figure 9: Seed budget and decision cost on frozen decisions. (A) Utility by updater-seed count for Direct CTV and debiased variants; dashed line is the post-hoc oracle. (B) Utility by decision cost for switching policies.

C.5 Debiased measurement

Debiased measurement combines a one-seed estimate $\hat{\tau}_i$ of candidate i, whose noise variance σ_i^2 the per-question spread estimates, with the pooled effect $\hat{\mu}_{-i}$ of the other candidates at the weight $\lambda_i = \kappa^2/(\kappa^2 + \sigma_i^2)$, where κ^2 is the spread of the estimates across candidates net of the mean noise variance, floored at 10^{-12}, and admits when the combination is positive.

Proposition 3 (Calibrated combination)**.** *Let the effects τ_i be exchangeable with mean μ and variance κ^2, and let $\hat{\tau}_i - \tau_i$ be mean-zero noise independent of τ_i, with variance σ_i^2. If $\kappa^2 + \sigma_i^2 > 0$, the weight $\lambda_i = \kappa^2/(\kappa^2 + \sigma_i^2)$ minimizes the mean squared error of $\lambda\hat{\tau}_i + (1-\lambda)\mu$ as an estimator of τ_i. The minimum is $\kappa^2\sigma_i^2/(\kappa^2 + \sigma_i^2)$, no larger than either variance and strictly smaller when both are positive.*

Proof. Write $e = \hat{\tau}_i - \tau_i$ and $d = \mu - \tau_i$, with variances $v_i = \sigma_i^2$ and $s^2 = \kappa^2$. The noise e is independent of τ_i, so e and d are independent with mean zero. The error of the combination is $\lambda e + (1-\lambda)d$, and $\mathbb{E}[(\lambda e + (1-\lambda)d)^2] = \lambda^2 v_i + (1-\lambda)^2 s^2 = s^2 - 2\lambda s^2 + \lambda^2(s^2 + v_i)$. For any a, c and $b > 0$, $a - 2\lambda c + \lambda^2 b - (a - c^2/b) = (\lambda b - c)^2/b \geq 0$, with equality at $\lambda = c/b$. Taking $a = c = s^2$ and $b = s^2 + v_i$ gives the minimizer $\lambda_i = s^2/(s^2 + v_i)$ and the minimum $s^2 - s^4/(s^2 + v_i) = s^2 v_i/(s^2 + v_i)$, which is below v_i and below s^2 whenever both are positive. □

The implementation substitutes $\hat{\mu}_{-i}$ for μ. The known-mean guarantee in Proposition 3 does not account for uncertainty in that estimate or for shared noise.

Seed budget on the frozen decisions. Figure 9 replays Direct CTV and debiased measurement at every updater-seed budget on the frozen decisions, averaging over seed subsets. At one seed, 24 continuation fits, debiased measurement realizes $+0.064$ points against $+0.036$ for Direct CTV, with 13.0 harmful admissions against 15.8. From two seeds on, Direct CTV leads and reaches $+0.174$ points at four seeds against $+0.145$. A prediction from the seven dataset-level scores lowers debiased utility at every budget, and the post-hoc oracle realizes $+0.27$ points, so Direct CTV at four seeds recovers 63% of the attainable utility on these decisions.

C.6 A model for the dose transition

Let $N = 16$ and denote the mean and standard deviation of realized candidate-minus-incumbent effects at dose D by $\mu(D)$ and $w(D)$. The following calculation assumes (A1) independent $\mathcal{N}(\mu(D), w(D)^2)$ effects across decisions, where w includes both dataset variation and updater noise; (A2) independent effects at the two receiving states of a dataset; and (A3) affine extrapolations of μ and w in $\log_2 D$ from the two measured doses, with $w > 0$. Set $z = \mu/w$, let ϕ, Φ be the standard normal density and CDF, and define $\psi(a) = \phi(a) - a\Phi(-a)$.

Proposition 2 gives the count, sign-flip probability, and oracle-gain bounds under (A1)–(A2).

One expected harmful decision occurs at $z_N = \Phi^{-1}(1 - 1/N)$. Write $\ell = \log_2(D/D_1)$ for the smaller measured dose and fit $\mu = \mu_1 + \beta\ell$ and $w = w_1 + \gamma\ell$. If $\beta - z_N\gamma \neq 0$ and the solution lies beyond the measured doses with $w > 0$, (A3) gives $D^\star = D_1 2^{(z_N w_1 - \mu_1)/(\beta - z_N\gamma)}$. For these sixteen decisions, the fitted crossing is approximately 637,000 tokens, beyond the larger measured dose. It has no uncertainty interval and is a model extrapolation, not an observed threshold. The fitted normal's counts and gains use the same sixteen effects at each dose as the observed summaries, so their agreement is an internal fit rather than independent validation. Effects at the two states are positively correlated at both doses (0.45 and 0.62); this departs from (A2) and can lower the sign-flip probability relative to the independent-state prediction.

C.7 Complete-state restoration

For AdamW, the first moment m, second moment v, and step clock t determine the bias-corrected update as well as the parameters θ. The paired continuation therefore restores $x = (\theta, m, v, t, \omega)$, including framework random state ω, before running either token- and step-matched pack. If two branch states have matching tensor shapes and optimizer groups, a common clock, finite values, and nonnegative second moments, their coordinatewise midpoint $(\frac{1}{2}(\theta_U + \theta_S), \frac{1}{2}(m_U + m_S), \frac{1}{2}(v_U + v_S), t)$ is an admissible AdamW state: averaging preserves shape and finiteness, preserves nonnegativity of v, and leaves the bias-correction clock unchanged. This construction supplies one shared reference, while exact reuse of its serialized state supplies the identity in Theorem 2.

Theorem 2 (Same-action identity under deterministic continuation)**.** *Fix one materialized complete state, its random state, the model configuration, and evaluation inputs. Write $F(ha)$ for the held-out result under history label h and next action a. Suppose each history label $h \in \{U, S\}$ installs this same state and random state before the next block, and that the installed state, next action $a \in \{U, S\}$, its packed data, and random state determine training and evaluation deterministically. Then $F(US) = F(SS)$ and $F(UU) = F(SU)$; hence $\tau_U = F(US) - F(UU) = F(SS) - F(SU) = \tau_S$ and the history interaction $I = \tau_U - \tau_S$ is zero.*

Proof. For each fixed next action, both history labels provide identical complete inputs to the deterministic continuation and evaluation, so they produce identical outcomes. Subtraction gives the two effect identities. □

The implementation checks restored state and random-state digests and replays same-action continuations, including final adapter digests and per-question answers. A failed replay invalidates the measurement. In the method-held-out Qwen3.5-2B control, complete-state reuse matches all 768 of 768 paired responses and gives $I = 0$. Restoring parameters or optimizer state alone leaves signed same-action accuracy differences of -1.0417 and -0.5208 points, respectively. Independently constructed complete states yield only 621 of 768 identical responses. Thus the observed identity depends on installing one saved state, rather than merely using the same state-construction rule.

C.8 Reference construction and state intervention

The midpoint above is symmetric in the two first-block histories; the branch states provide two additional references. Each reference can continue on either next dataset, and its held-out value decomposes as $V_r = B_r + G_r$, where B_r retains the incumbent and G_r is the gain from the selected next dataset. Midpoint symmetry alone does not guarantee a better V_r. The parameter-only and optimizer-only restorations above isolate the need to retain the full AdamW state; Figure 2(B) reports the moment-reset order contrast.

C.9 Reference-choice check

The reference-choice study crosses Qwen3.5-2B and OLMo3-7B with panels B and D and three updater seeds, giving twelve core blocks; six Qwen3.8-27B blocks form a separate scale group. Each reference uses eight steps, chooses the next dataset on 256 questions, and evaluates it on 512 disjoint questions. The 108 primary continuations and nine same-action replicas give 117 trained students. The panels were exposed during method development.

Across the twelve core blocks, the midpoint gives 17.2852% selected accuracy against 17.6921% for the equal anchor mixture. The difference is -0.4069 points, comprising -0.3499 retained-baseline points and -0.0570 selection-gain points. Only three of twelve block effects are positive; midpoint symmetry does not establish a beneficial reference choice.

D Evaluation Scope

D.1 Policy inputs and comparator scope

Table 1 groups the policy inventory by what can change its decision between receiving states. Dataset-only inputs cover admit-all, teacher accuracy (Li et al., 2023), execution correctness, near-duplicate removal, and convex valuation (Zeng et al., 2026). Scores that read the dataset through current parameters cover perplexity, student likelihood, the receiving-model judge, RSR, forward learnability (Yu et al., 2026), TracIn (Pruthi et al., 2020), LESS (Xia et al., 2024), TSDS (Liu et al., 2024), Data-DPO (Sun et al., 2026), and PRISM (Lin et al., 2026). Proxy-model inputs cover the 1.7B transport student, the one-step micro-model assay, MATES (Yu et al., 2024), and RHO-LOSS (Mindermann et al., 2022). Direct CTV and debiased measurement read the complete receiving state. The inventory includes unevaluated policies; Table 6 reports only the tested decision policies and random selection.

The scoring baselines in the frozen comparisons adapt cited methods to whole-dataset admission. The cross-state gradient cosine is a base-state candidate-pool proxy inspired by LESS; its ranking differs from the official warmup and checkpoint influence procedure. A subsequent same-pool study includes a BIRD adaptation of that procedure. The frozen benchmark and the new comparison retain distinct comparator sets; an AdamW-preconditioned short-horizon influence comparator remains untested (Deng et al., 2026).

The same-pool study uses Qwen3-4B, 3004 candidate traces and 1502 selected traces per arm. The four arms are official LESS adapted to BIRD, the earlier gradient-cosine proxy, uniform selection, and CTV family selection. The latter ranks traces by their family's frozen mean selection effect. Each arm trains at 65,536 tokens and 32 steps under three updater seeds, and reads the same 441 validation questions from six databases. Validation had been inspected before this comparison was proposed, so the results are exploratory. CTV family selection has utility -2.19 points relative to the incumbent. Its gaps to official LESS, the proxy and uniform selection are $+5.06$, $+5.22$ and -0.53, with one-sided 0.05 lower bounds $+1.23$, $+2.27$ and -3.85. The 10,000 shared bootstrap draws resample updater seeds and validation databases jointly across arms. Equal downstream budgets do not equalize selection cost: LESS includes warmup and gradient extraction, while CTV reuses 27 pilot continuations and their readouts, charged once across the three downstream seeds.

D.2 Additional selection comparisons

In separate native-schedule comparisons on four producer populations, LARK's more favourable top-two variant gains 0.79 execution-accuracy points over random-two, while DIVE loses 0.74 against uniform-source selection and TracInCP loses 0.28 against random-four (Yu et al., 2026; Nyakairu & Liu, 2026; Pruthi et al., 2020). Each arm and control answers the same 441 validation queries per population; paired tests do not reject zero effect ($P = 0.38, 0.44, 0.79$ respectively). These comparisons preserve each selector's eligibility and training schedule, but do not reproduce its original benchmark. They therefore neither establish selector gains here nor refute gains in the original settings.

At a common 524,288-token, 32-update budget, a separate grid compares one-checkpoint gradient dot product, candidate loss, and gradient cosine with uniform random and length- and format-matched null packs. On the custom split no rule, null, or Menu CTV exceeds uniform random by more than 0.1 validation points. This grid tests scores within one candidate pool; it is distinct from the native-schedule selector comparisons above and the dataset-level decisions in Table 6. On official BIRD dev, an exploratory three-seed batch favours Menu CTV, but its fixed-pack reference was selected after seeing those seeds. Repeated scoring of the uniform-random control gives 24.90, 25.68, and 26.99 percent rather than the batch's 27.12, 27.84, and 28.16 percent. No seed in the registered fresh-seed replication had all ten packs complete at the analysis cutoff. The official-dev relative gain is therefore unresolved.

D.3 Evaluator and serving boundaries

The recorded SQL extraction and fenced-only sensitivity for the frozen benchmark are specified in Appendix C.4. Both tracks retain positive Direct CTV utility, but the candidate data can change whether the student writes a code block; the extraction rule is part of the measured utility. Likewise, Direct CTV estimates a candidate-versus-incumbent continuation under a specified parser, evaluator, and serving configuration. An estimate need not transfer when one of these changes.

Three checks show why those qualifications matter. First, a Gemma replay of the reference study yields 4 versus 3,182 successes when two SQL extractors score the same 9,894 saved responses; this partially observed replay does not give a complete paired effect. Second, a transfer of BIRD-selected reference policies to LiveSQLBench (BIRD Team, 2024) encounters an official evaluator that accepts `SELECT 1` on 15 of 90 management tasks. A local repair resolves 12 tasks while three semantic branches remain unresolved; replaying alternate SQL fences also changes success counts, especially for Qwen3.5. The LiveSQL comparison is conditional on its declared parser and evaluator and cannot be read as a BIRD result. Third, in four WebShop paired continuations (Yao et al., 2022), complete transcripts match in only 31 of 320 cases, although reward and success agree in 307. The declared exact-response identity fails in interactive serving, leaving the paired data effect unavailable under that criterion. These are different experiments with different units and evaluators; none supplies a replacement estimate for another.