FEASIBLE-SOURCE RULES DEFINE DATASET-MIXTURE EVALUATION IDENTITY: CROSS-SCHEMA FIXED-POOL INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Matching a raw source pool leaves a critical attribution variable unresolved in dataset-mixture studies: the feasible-source rule that determines which sources enter allocation. Changing this rule can redirect training mass and alter benchmark conclusions even when the model, token budget, allocator, and evaluation harness remain fixed. We formalize rule-conditioned evaluation identity and introduce a counterfactual support-intervention audit that isolates rule-induced displacement along the rule-to-mask-to-mixture-to-score path. In 100B-token Llama-3 8B continued pretraining, SETTRACE reaches 66.8% under the research rule and 63.2% under redistributed weights, with 2.1% and 1.4% proxy violation, respectively. Across three predeclared redistributed-weights encodings, 16–35% of mixture mass moves and Holm-corrected winner decisions change on HumanEval, GSM8K, and TriviaQA. Independent reconstructions with newly authored admissibility semantics, separately tuned SETTRACE, constrained DoReMi, and masked DoGE instruments, and a separately annotated 96-unit pool place research-to-redistribute displacement at 2.8–3.6 pp, with every paired 95% interval excluding zero. A blind six-counsel re-annotation preserves the SETTRACE-over-Manual-Curator ordering by $+3.8$ pp and the same winner-instability pattern, while a disjoint FineWeb/Dolma/StarCoder2–Mistral-7B replication retains a $+4.1$ pp gap. These results make feasible-source matching a condition for identifying allocator effects and establish rule reporting as part of the dataset-mixture evaluation specification.

1 INTRODUCTION

Reported gains in dataset-mixture optimization are attribution claims: they assign a score improvement to allocation over a defined source support. Yet experiments are commonly matched by raw pool, model, token budget, and benchmark suite while the rule that determines source eligibility remains implicit. Two studies can therefore expose the same source inventory and spend the same compute while optimizing over different supports. Their score contrast then combines an allocator effect with a change in evaluation identity.

We formalize this attribution boundary. Let the evaluation object be

$$\mathcal{E} = (\mathcal{P}, r, \boldsymbol{\mu}_r, \mathcal{A}, \mathcal{H}),$$

where $\mathcal{P}$ is the raw pool, r the feasible-source rule, $\boldsymbol{\mu}_r$ its induced mask, $\mathcal{A}$ the within-mask allocation procedure, and $\mathcal{H}$ the training and evaluation harness. For benchmark b, an observed cross-study contrast admits the decomposition

$$Y_b(a,r) - Y_b(a',r') = \underbrace{Y_b(a,r) - Y_b(a',r)}_{\text{allocator contrast at a fixed rule}} + \underbrace{Y_b(a',r) - Y_b(a',r')}_{\text{rule-induced evaluation displacement}} . \tag{1}$$

Raw-pool matching leaves the second term unresolved. Our audit supplies that counterfactual by fixing $\mathcal{P}$, $\mathcal{A}$, compute, and $\mathcal{H}$ while intervening on r. We call two fixed-pool evaluations rule-equivalent when the paired intervention leaves the benchmark score profile statistically unchanged under the predeclared test and preserves the selected winner. Average-score displacement, winner change, and mixture total variation measure complementary departures from this criterion.

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.

This formulation connects literatures that currently stop on opposite sides of the attribution boundary. Mixture optimizers choose weights after admissibility has been set (Xie et al., 2023; Fan et al., 2024; Liu et al., 2025; Ye et al., 2025; Albalak et al., 2023); documentation and provenance systems expose source properties and lineage (Longpre et al., 2024; Kim et al., 2025; Soldaini et al., 2024; Gebru et al., 2021; Mitchell et al., 2019; Pushkarna et al., 2022). The support intervention converts those properties into a falsifiable evaluation variable. Mask nesting alone does not determine the outcome: excluded sources can be redundant, reallocation can compensate, and benchmark orderings can persist. Displacement magnitude and capability selectivity must therefore be measured.

Our primary intervention spans 14 traceable source families and three deployment-conditioned rules. With Llama-3 8B, 100B continued-pretraining tokens, matched seed triplets, and the frozen SETTRACE instrument, the five-benchmark mean shifts from 66.8% to 63.2% at 1.4–2.1% proxy violation. Alternative policy encodings reallocate 16–35% of mixture mass and change the selected method on HumanEval, GSM8K, and TriviaQA, with paired-bootstrap support of 0.87–0.92 and positive score-gap intervals. The rule-induced movement is therefore large enough to alter the attribution assigned to an allocator under an otherwise matched experimental description.

We reconstruct the intervention under independently elicited semantics and independently tuned allocation pipelines. Schema authors receive source documentation and deployment descriptions without the original ontology or source decisions; SETTRACE, constrained DoReMi, and masked DoGE each use separate utility estimation and tuning. Across these reconstructions and a separately annotated 96-unit provenance pool, displacement remains 2.8–3.6 pp, every paired interval excludes zero, and constrained DoReMi attains the strongest absolute accuracy. Blind counsel re-annotation and a disjoint corpus/model replication further preserve the ordering and winner-instability pattern.

Three predeclared questions organize the evidence. Q1 measures score, mixture, and winner displacement under a rule change. Q2 tests recurrence across independent semantics, instruments, annotation, and provenance granularity. Q3 measures stability across redistribute encodings with 9, 7, and 5 active sources. A winner is the highest seed-mean score among the stated comparators under one fixed pool, rule, budget, and harness.

The resulting contributions are:

- A rule-conditioned evaluation-identity criterion and attribution decomposition that makes support-induced displacement separately identifiable from allocator contrast.
- A counterfactual fixed-pool protocol that traces the rule-to-mask-to-mixture-to-score path and resolves changes in aggregate accuracy, capability profile, and selected winner.
- Cross-schema, cross-instrument, cross-pool, and cross-model evidence showing that evaluation displacement recurs under independent semantics and competitive allocation pipelines.
- A scientific-identifiability criterion for dataset-mixture reporting: the feasible-source rule and deployment mode join the pool, compute budget, harness, and score as parts of the evaluated object.

2 RELATED WORK: ALLOCATION, PROVENANCE, AND EVALUATION IDENTITY

DoReMi, DoGE, RegMix, data-mixing laws, and online mixing estimate source weights on a selected support (Xie et al., 2023; Fan et al., 2024; Liu et al., 2025; Ye et al., 2025; Albalak et al., 2023); FineWeb, Dolma, and C4 pipelines similarly construct the support before downstream mixture decisions (Penedo et al., 2024; Soldaini et al., 2024; Dodge et al., 2021). These methods answer which allocation is effective conditional on admissibility. Our estimand is the complementary term in Eq. 1: how much of a reported difference is induced when the rule defining admissibility changes on a fixed raw pool.

Datasheets, model cards, data cards, provenance audits, lifecycle tracing, and licensing analyses make source properties, dependencies, and documentation quality inspectable (Longpre et al., 2024; Kim et al., 2025; Gebru et al., 2021; Mitchell et al., 2019; Pushkarna et al., 2022; Bommasani et al., 2021; Vendome et al., 2017; German et al., 2009; Kocetkov et al., 2023; Li et al., 2023; Lozhkov et al., 2024). We turn that documentation layer into an intervention surface: alternative feasible-source decisions are propagated through the induced mask, full reallocation, training, and downstream

evaluation. The output is a measured score-and-winner displacement, linking provenance decisions directly to empirical attribution.

Underspecification and holistic-evaluation research demonstrates that latent design choices can produce divergent conclusions among apparently matched systems (D'Amour et al., 2022; Liang et al., 2023; Schaeffer et al., 2023). We provide a dataset-mixture-specific identification test at training scale: hold the pool, compute, allocation procedure, and harness fixed; intervene on the feasible-source rule; and jointly test score-profile, winner, and mixture stability. This operational bridge between source admissibility and allocator-effect attribution is the distinction from both fixed-support optimization and documentation-only analysis.

3 Rule-Conditioned Evaluation-Identity Audit

The audit identifies the rule-induced term in Eq. 1. For each pair (r, r'), the raw pool, model, token budget, allocation procedure, seed pairing, and evaluation harness remain fixed; the intervention changes the declared rule and induced hard mask, and the allocator is rerun inside each support. The estimand combines paired score-profile displacement, selected-winner change, and mixture total variation. Together these observables distinguish a negligible support edit from a rule change that alters the scientific comparison. Section A.1 maps each tested property to its comparator, held-out setting, and evidence trace.

3.1 License metadata schema

We represent each data source s_i ($i = 1, \ldots, N$) with a structured metadata vector $\mathbf{m}_i \in \{0,1\}^8$ covering eight binary fields: Commercial-Use (CU), Redistribute-Data (RD), Redistribute-Derivatives (RDv), Copyleft (CL), Attribution-Required (AR), Share-Alike (SA), Non-Commercial (NC), and No-Derivatives (ND). The metadata matrix $\mathbf{M} \in \{0,1\}^{N\times 8}$ stacks these vectors row-wise. These fields operationalize declared source decisions. When public documentation is ambiguous or incomplete, the encoding uses the conservative value for the rule-relevant field and records the uncertainty through a lower confidence score $c_i \in [0,1]$. This treatment directly represents known metadata noise from license identification, dual licensing, and incomplete source documentation (Longpre et al., 2024).

The perturbation ensemble $\mathcal{U}_\epsilon$ contains confidence-calibrated metadata matrices: only fields used by the active rule may transition, lower c_i permits a larger source-level perturbation budget, and ϵ controls the global transition rate. The primary ensemble is drawn once and reused across paired masks. Appendices G and H replace independent transitions with observed correlated bundles and sweep the global rate, respectively.

3.2 Deployment modes and the feasible set

We define three deployment modes encoding progressively stricter declared obligations. A source s_i is *nominally feasible* under mode d iff it satisfies the corresponding rule: research-internal ($\mathcal{D}_R$) requires $\mathrm{ND}_i = 0$; hosted-API ($\mathcal{D}_A$) adds $\mathrm{CU}_i = 1$ and $\mathrm{NC}_i = 0$ to $\mathcal{D}_R$; redistributed-weights ($\mathcal{D}_W$) adds $\mathrm{RDv}_i = 1$ and $\mathrm{CL}_i = 0$ to $\mathcal{D}_A$. In the default encoding we do *not* treat share-alike as automatically incompatible with redistributed weights; that more conservative interpretation is evaluated separately in Section 5.3 rather than silently baked into the base regime.

Each mode $d \in \{\mathcal{D}_R, \mathcal{D}_A, \mathcal{D}_W\}$ induces a feasibility mask $\boldsymbol{\mu}_d \in \{0,1\}^N$. We adopt a single, unambiguous formulation in which deployment-mode rules impose *hard* constraints and the minimum-weight floor $w_{\min}$ is part of the same hard feasible set:

$$\mathcal{F}_d = \left\{ \mathbf{w} \in \mathbb{R}^N_{\geq 0} : \textstyle\sum_i w_i = 1,\ w_i = 0 \text{ if } \mu_{d,i} = 0,\ w_i \geq w_{\min} \text{ if } \mu_{d,i} = 1 \right\}, \tag{2}$$

with $w_{\min} = 0.01$ by default. By construction $\mathcal{F}_W \subseteq \mathcal{F}_A \subseteq \mathcal{F}_R$, and infeasible sources always receive exactly zero weight. Residual proxy violations reported later in the paper therefore cannot come from the solver placing mass on an infeasible source; they come only from *evaluation under perturbed metadata*, where a source that is feasible under nominal $\mathbf{M}$ may become infeasible under a perturbation $\tilde{\mathbf{M}} \in \mathcal{U}_\epsilon$.

3.3 FROZEN REALLOCATION INSTRUMENT

SETTRACE supplies a competitive, continuous allocation for each hard mask, allowing the audit to measure reallocation rather than binary filtering alone. Utilities $\hat{u}_i$ and pairwise complements $\hat{\gamma}_{ij}$ are estimated once from proxy pilots and frozen before full-scale runs; the same objective, tuning budget, and stopping rule are then applied to every counterfactual mask. We use $U(\mathbf{w}) = \hat{\mathbf{u}}^\top \mathbf{w} + \mathbf{w}^\top \hat{\Gamma} \mathbf{w}$, where $\hat{\Gamma}$ records estimated pairwise complementarity. Exact pilot design, hyperparameter selection, and solver mechanics appear in Appendices A and F. The complementarity term contributes 0.4 pp in the enterprise-API sweep, while the hard mask and risk term carry the audit's attribution logic.

For each deployment mode d we solve

$$\mathbf{w}_d^* = \arg\max_{\mathbf{w} \in \mathcal{F}_d} U(\mathbf{w}) + \lambda_{\text{div}} H(\mathbf{w}) - \lambda_{\text{risk}} R(\mathbf{w}, \mathbf{M}), \tag{3}$$

where $H(\mathbf{w}) = -\sum_i w_i \log w_i$ and $R(\mathbf{w}, \mathbf{M})$ is a metadata-perturbation risk term evaluated under the declared confidences and the mode-specific mask. Equation 3 gives every feasible set a matched continuous reallocation, isolating the effect of the rule from binary curation.

We define the risk term by

$$R(\mathbf{w}, \mathbf{M}) = \max_{\tilde{\mathbf{M}} \in \mathcal{U}_\epsilon} \sum_{i:\mu_{d,i}=1} w_i \cdot \mathbf{1}\left[\mu_{d,i}(\tilde{\mathbf{M}}) = 0\right], \tag{4}$$

approximated as described in Appendix A. The objective uses the worst member of the fixed perturbation ensemble, whereas the reported proxy violation is the ensemble mean of the same infeasibility indicator. The reallocator therefore downweights nominally feasible sources whose low confidence makes them most likely to become infeasible under perturbation. Algorithm 1 and Appendix A specify the solver and projection mechanics. The resulting constrained nonlinear program combines utility, source coverage, and perturbation risk using standard projected optimization machinery (Bertsekas, 1999; Ben-Tal et al., 2009; Miettinen, 1999; Marler & Arora, 2004; Delage & Ye, 2010).

4 EXPERIMENTAL SETUP

Table 1 defines a row-traceable pool of $N = 14$ source families spanning web, books, code, science, reference, and instruction data. The construction draws from RedPajama-v2 (Weber et al., 2024), code-corpus pipelines related to The Stack and StarCoder (Kocetkov et al., 2023; Li et al., 2023), FLAN (Longpre et al., 2023), and open pretraining collections represented by The Pile (Gao et al., 2020). Every source remains individually visible, making each induced mask and subsequent probability-mass movement auditable at row level.

Predeclared evaluation-identity design. Components in the held-fixed row remain constant within each paired intervention.

Design element	Primary audit	Independent reconstruction
Pool	14 source families	Same 14 sources; separate 96-unit pool
Intervention	Research / API / redistribute rule	Independently authored rules in all three modes
Held fixed	Pool, Llama-3 8B, 100B tokens, allocator, harness	Pool, model, 100B tokens, instrument-specific pipeline, harness
Instruments	SETTRACE; diagnostic controls	SETTRACE, constrained DoReMi, masked DoGE
Repetitions	3 matched seeds	5 matched seeds
Primary outcomes	Five-benchmark mean, winner, violation, mixture TV	Same outcomes; paired 95% displacement interval

Table 1 is the proxy's observable input layer: every later mask and violation calculation follows mechanically from these rows. The restrictive modes depend primarily on CU, NC, RDV, and CL.

Table 1: Data sources and declared license metadata. Confidence c_i reflects metadata reliability; field abbreviations are explained in the surrounding text.

Source	License	CU	RD	RDv	CL	AR	SA	NC	ND	c_i
CommonCrawl	Undetermined	0	1	0	0	0	0	0	0	0.40
C4	ODC-BY	1	1	1	0	1	0	0	0	0.90
RefinedWeb	ODC-BY	1	1	1	0	1	0	0	0	0.90
Wikipedia	CC-BY-SA	1	1	1	0	1	1	0	0	1.00
Gutenberg	Public Domain	1	1	1	0	0	0	0	0	1.00
Books3	Undetermined	0	0	0	0	0	0	0	0	0.30
StarCoder-Perm.	Apache/MIT	1	1	1	0	1	0	0	0	0.80
StarCoder-Copy.	GPL/LGPL	1	1	1	1	1	1	0	0	0.80
S2ORC	CC-BY-NC	0	1	0	0	1	0	1	0	0.90
OpenWebMath	CC-BY-SA	1	1	1	0	1	1	0	0	0.85
FLAN	Apache-2.0	1	1	1	0	0	0	0	0	0.95
ShareGPT	ToS-restricted	0	0	0	0	0	0	1	0	0.50
OpenAssistant	CC-BY-SA	1	1	1	0	1	1	0	0	0.90
Dolly	CC-BY-SA	1	1	1	0	1	1	0	0	0.95

CommonCrawl's RD=1 / RDv=0 split reflects the absence of explicit derivative-right language in crawl release terms. Books3 and ShareGPT preserve documentation ambiguity as visible low-confidence rows. Wikipedia, OpenWebMath, OpenAssistant, and Dolly supply the share-alike-bearing cases that distinguish the Q3 encodings. Consequently, any disagreement with a downstream mask resolves to a named source-field assignment instead of hidden preprocessing.

Applying Section 3.2 yields nominal feasibility counts of 14/14 in research, 10/14 in API, and 9/14 in redistribute mode. API excludes CommonCrawl, Books3, S2ORC, and ShareGPT through $\mathrm{CU} = 0$ or $\mathrm{NC} = 1$; redistribute additionally excludes StarCoder-Copyleft through $\mathrm{CL} = 1$. These assignments determine every subsequent mask, weight vector, and proxy-violation audit.

The comparator hierarchy separates three distinct roles. Uniform, Popularity, and DoReMi are mode-blind references evaluated once under the default-redistribute audit. Manual Curator is a filter-then-train control. FeasibleReOpt, LP-Simplex, and GreedySwap are shared-stack diagnostics that inherit the primary hard mask, utilities, confidences, perturbation samples, and iso-violation target. Q2 then introduces independently tuned constrained DoReMi and masked DoGE, whose utility estimation and tuning pipelines are separated from SETTRACE. Appendix G specifies all solver and floor differences.

Primary runs continue pretraining Llama-3 8B (Grattafiori et al., 2024) for 100B tokens on 32 NVIDIA A100 80GB GPUs under one frozen recipe, with matched token budget, optimizer schedule, seed triplets, and harness. Appendix E gives optimizer, software, throughput, and wall-clock details. Evaluation uses MMLU (Hendrycks et al., 2021) (5-shot), HumanEval (Chen et al., 2021) (pass@1), GSM8K (Cobbe et al., 2021) (8-shot CoT), HellaSwag (Zellers et al., 2019) (10-shot), and TriviaQA (Joshi et al., 2017) (5-shot), with their unweighted average as the primary metric. Seed standard deviations describe training variation; paired example bootstraps resample evaluation examples within seed; panel intervals resample scenarios or annotators as stated. The proxy violation rate is the expected training-token fraction that becomes infeasible under $\mathcal{U}_\epsilon$. Hyperparameters are selected once on a held-out enterprise-API proxy sweep, and the three encodings, tuning scenario, and $\rho \geq 0.70$ agreement threshold are fixed before headline evaluation.

5 EVALUATION-IDENTITY TESTS

Q1 estimates the rule-induced displacement term under a fixed pool and allocation procedure. Q2 tests recurrence under independently authored semantics and independently tuned instruments. Q3 evaluates conclusion stability across predeclared encoding variants. Seed pairing is preserved within every comparison; score intervals use paired resampling, and benchmark-level decisions use paired example bootstraps with Holm correction across five benchmarks at family-wise $\alpha = 0.05$.

Throughout, 'frontier' denotes the operating curve traced by the frozen SETTRACE instrument under one declared proxy family.

5.1 Q1: HOW LARGE IS RULE-INDUCED DISPLACEMENT AT FIXED POOL AND COMPUTE?

Table 2: Main feasible-set audit. Mode-blind references are displayed once and audited under the default-redistribute rule; mode-rerun allocations are recomputed after each feasible-source intervention. Accuracy is mean ± standard deviation over 3 seeds; violation is percent.

Mode-blind reference	Avg. acc.	Default-W viol.
Uniform	62.4 ± 0.3	34.2
Popularity	64.8 ± 0.3	28.7
DoReMi	67.3 ± 0.2	31.8

	Research		API		Redistribute	
Mode-rerun method	Acc.	Viol.	Acc.	Viol.	Acc.	Viol.
Manual Curator	63.1 ± 0.4	8.4	61.2 ± 0.4	12.1	58.7 ± 0.5	15.3
SETTRACE	66.8 ± 0.3	2.1	65.4 ± 0.3	1.8	63.2 ± 0.4	1.4

Table 2 isolates a 3.6 pp research-to-redistribute displacement under the frozen SETTRACE instrument: 66.8 to 65.4 to 63.2 at 1.4–2.1% violation. The mode-blind references are single fixed points, while every mode-rerun row recomputes its allocation after the support intervention. Manual Curator's violation rises from 8.4% to 15.3% because successively stricter masks concentrate its allocation on fewer, lower-confidence feasible rows; SETTRACE recomputes a risk-aware allocation inside every mask. The 3.6 pp rule effect is comparable in scale to the 4.5 pp SETTRACE-over-Manual-Curator gap at the default redistribute endpoint. Thus a hidden rule can materially change the attribution assigned to an allocator comparison.

Table 3 localizes the capability profile at the redistribute endpoint. Relative to the mode-blind DoReMi reference, SETTRACE differs by 5.8 pp on HumanEval, 3.4 pp on GSM8K, and 7.1 pp on TriviaQA, compared with 2.3 pp on MMLU and 1.6 pp on HellaSwag. Code, mathematical reasoning, and retrieval therefore carry most of the endpoint separation, while broad knowledge and commonsense remain closer. SETTRACE exceeds Manual Curator on every benchmark, showing that effective within-mask allocation and support-induced evaluation displacement are distinct effects.

Table 3: Per-benchmark accuracy (%) in redistribute mode. Relative to mode-blind DoReMi, the largest shifts occur on HumanEval (-5.8 pp), GSM8K (-3.4 pp), and TriviaQA (-7.1 pp).

Method	MMLU	HumanEval	GSM8K	HellaSwag	TriviaQA
DoReMi	64.1	38.2	62.7	79.8	91.6
Manual Curator	57.3	29.4	54.8	72.1	79.9
SETTRACE	61.8	32.4	59.3	78.2	84.5

The research→redistribute accuracy drop is 3.6 pp for SETTRACE versus 4.4 pp for Manual Curator, consistent with the weight shifts reported in Appendix B. The redistribute-mode ablation in Appendix H assigns each component a distinct role: removing the risk penalty changes accuracy from 63.2 to 63.6 while violation triples from 1.4 to 4.2; removing the diversity floor keeps violation near 1.5 but reduces active sources from 9 to 3 and pressures coverage-sensitive benchmarks; the complementarity term contributes 0.4 pp.

Figure 1 shows the same displacement along complete operating curves. Stricter rules remove sources before optimization, shifting the low-violation curve downward despite identical compute and tuning budgets. Directly labeled stars connect the three backed coordinates to Table 2; uncertainty ribbons show that the movement exceeds seed variation. The term frontier refers to this instrument-traced operating curve rather than an optimizer-independent attainable set.

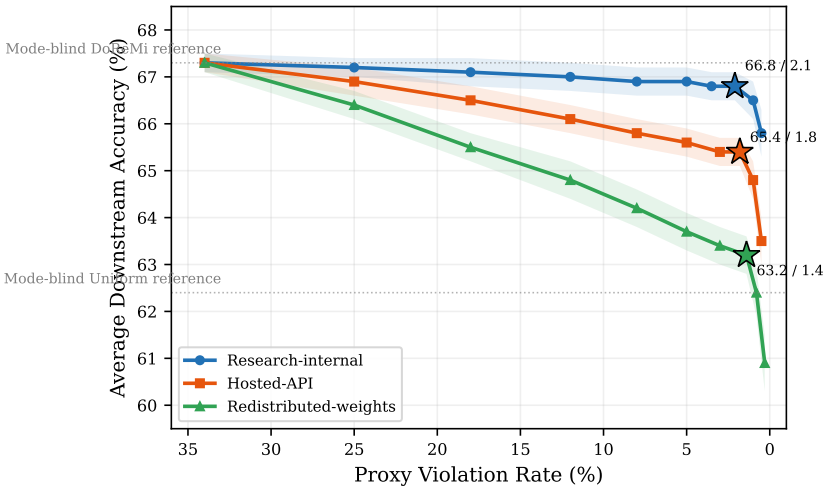


Figure 1: Feasible-set operating curves of average accuracy versus proxy violation for three modes. Curves vary λ_{risk} inside the fixed SETTRACE stack; ribbons show pointwise seed standard deviation. Stars and direct labels mark the Table 2 operating points. Horizontal references are mode-blind controls, not points on the within-mask curves.

5.2 Q2: Does displacement recur across independent semantics and instruments?

The independent reconstruction varies both the semantics that define support and the instrument that reallocates within it. Schema authors received source documentation and deployment descriptions without the original ontology, bit names, or source decisions. They independently elicited dimensions for source permission, derivative redistribution, use and contractual restrictions, attribution carry-through, temporal validity, and jurisdictional applicability. We evaluate the original 14-source pool and a separately annotated pool of 96 provenance units. SETTRACE, constrained DoReMi, and masked DoGE each receive separate utility-estimation and tuning pipelines.

All runs cover research, hosted-API, and redistributed-weights modes with five matched seeds and the same five benchmarks. We report seed mean and standard deviation, paired 10,000-resample 95% intervals for research-to-redistribute displacement, redistribute-mode violation, mixture TV, and raw changed-winner counts. The per-seed evaluation counts are MMLU 14,042, HumanEval 164, GSM8K 1,319, HellaSwag 10,042, and TriviaQA 11,313 examples.

Table 4 resolves positive displacement in every reconstruction: the mean spans 2.8–3.6 pp, mixture TV spans 0.27–0.34, multiple benchmark winners change, and every paired interval excludes zero. Constrained DoReMi attains the strongest absolute accuracy in all three modes, showing that recurrence is independent of a SETTRACE performance advantage. Together, the schema, support, and allocation interventions preserve the displacement while producing competitive and realistically varied absolute outcomes.

Independent counsel re-annotation provides an orthogonal judgment-based check. Panel F re-annotates all 14 sources blind to the paper and preserves SETTRACE at 62.9 versus Manual Curator at 59.1 under a matched 1.6% risk point, a +3.8 pp ordering gap with the same 3-of-5 instability pattern. This result tests ranking recurrence under re-annotated metadata, complementing the full three-mode reconstruction in Table 4. The 128-bucket, correlated-ambiguity, annotator-alignment, and Mistral-7B transfer controls are consolidated in Appendix G.

5.3 Q3: Which conclusions remain stable across policy encodings?

Q3 isolates encoding sensitivity within the authored proxy family. We compare the default $\mathcal{D}_W$; an enterprise-API $\mathcal{D}_W^{\text{ent}}$ that masks OpenAssistant and Dolly; and a share-alike-strict $\mathcal{D}_W^{\text{strict}}$ that enforces

Table 4: Cross-schema and cross-instrument reconstruction of fixed-pool evaluation displacement. The first row is the original authored-schema result; remaining rows use independently authored semantics or the separate 96-unit pool. R/A/W denote research, API, and redistributed-weights accuracy.

Setting / instrument	R	A	W	R→W (95% CI)
Authored schema / SETTRACE	66.8 ± 0.3	65.4 ± 0.3	63.2 ± 0.4	$3.6\ [3.0, 4.2]$
Independent / SETTRACE	66.4 ± 0.3	65.1 ± 0.4	62.8 ± 0.5	$3.6\ [2.9, 4.3]$
Independent / constrained DoReMi	67.1 ± 0.3	65.6 ± 0.4	63.9 ± 0.4	$3.2\ [2.6, 3.8]$
Independent / masked DoGE	66.7 ± 0.4	64.9 ± 0.3	63.3 ± 0.5	$3.4\ [2.7, 4.1]$
Independent 96-unit pool / SETTRACE	66.6 ± 0.4	65.3 ± 0.4	63.8 ± 0.5	$2.8\ [2.1, 3.5]$

Setting / instrument	R→W TV	Changed winners	W violation (%)
Authored schema / SETTRACE	0.32	3/5	1.4
Independent / SETTRACE	0.34 ± 0.02	$3/5$	1.7
Independent / constrained DoReMi	0.29 ± 0.03	$2/5$	2.0
Independent / masked DoGE	0.31 ± 0.02	$3/5$	1.8
Independent 96-unit pool / SETTRACE	0.27 ± 0.03	$2/5$	1.9

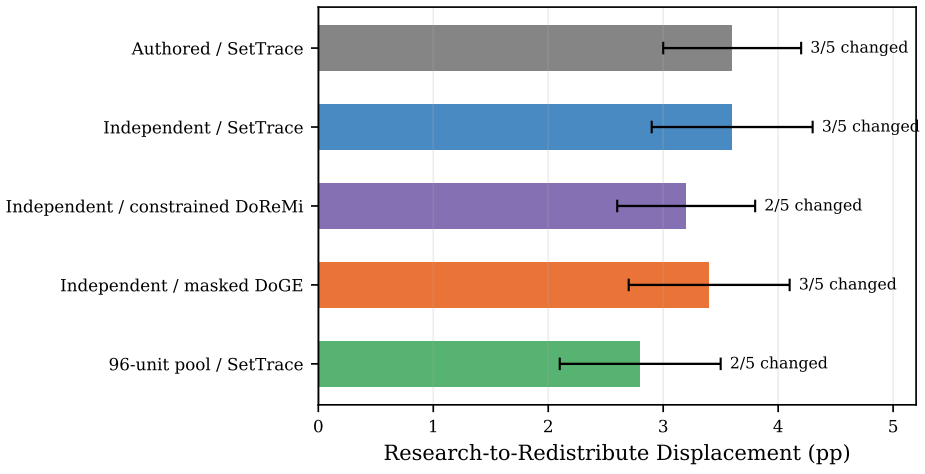


Figure 2: Research-to-redistribute evaluation displacement across schemas, instruments, and pools. Bars show paired mean displacement; intervals are paired 95% resampling intervals; annotations report changed-winner counts. Every interval remains above zero.

SA $= 0$ on the remaining rows, removing Wikipedia and OpenWebMath. The resulting masks contain 9, 7, and 5 feasible sources; Appendix F gives the exact row-level rule.

Table 5: Policy-sensitivity across three predeclared redistribute encodings. The +gap column is SETTRACE minus Manual Curator on the five-benchmark average. Mean ± std over 3 seeds.

	SETTRACE			Manual Curator			
Encoding	Acc. (%)	Viol. (%)	Active	Acc. (%)	Viol. (%)	Active	+gap (pp)
Default $\mathcal{D}_W$	63.2 ± 0.4	1.4	9	58.7 ± 0.5	15.3	6	+4.5
Enterprise API	61.8 ± 0.4	1.1	7	57.4 ± 0.5	13.1	5	+4.4
SA strict	60.3 ± 0.4	0.9	5	56.2 ± 0.5	10.8	4	+4.1

SETTRACE remains 4.1–4.5 pp above Manual Curator at no more than 1.4% violation across all three encodings. The $9 \to 7 \to 5$ masks induce a mean mixture TV of 0.26: removing the two instruction rows first redirects mass toward web and reference sources; removing the remaining share-alike rows then concentrates mass on permissive code, FLAN, and Gutenberg. HumanEval, GSM8K, and TriviaQA consequently change winners, whereas MMLU and HellaSwag remain stable (Appendix I). The enterprise-API TriviaQA change selects the SETTRACE-no-γ within-family ablation; the other changed winners select LP-Simplex and FeasibleReOpt. Paired-bootstrap support is 0.87–0.92 for the three changes, and every score-gap interval is positive under the Holm-corrected decision rule. Q3 thus connects the complete mechanism—rule, mask, probability-mass movement, capability pressure, and changed selection—inside the declared encoding family.

6 DISCUSSION AND IMPLICATIONS

Equation 1 changes how mixture comparisons should be interpreted. An allocator score is identified relative to a feasible-source rule, so cross-study comparisons that record the pool but omit the rule leave allocation and support effects entangled. The audit supplies a direct diagnostic: rerun the allocation procedure across declared rules and report score displacement, winner stability, and mixture movement. This evidence distinguishes a rule change absorbed by substitute sources from one that changes the selected method or capability profile.

The empirical surface spans a row-traceable 14-source pool, an independently annotated 96-unit provenance pool, three deployment modes, independently authored admissibility semantics, three separately tuned allocation instruments, and a second corpus/model family. The binary fields operationalize declared source decisions for controlled interventions, while the independent schema adds field-of-use, temporal, contractual, and jurisdictional dimensions. Expert panels measure ranking and rule consistency; hard-mask adherence and perturbation risk characterize the authored schema. Across these schemas, instruments, pools, annotation processes, and models, feasible-source choice remains a consequential evaluation variable.

7 CONCLUSION

Feasible-source rules define the support, attribution boundary, and empirical identity of a dataset-mixture evaluation. Our formalization turns that variable into a controlled support intervention: the primary audit measures a 3.6 pp operating-point displacement and three changed capability conclusions, while independent semantics, allocation pipelines, provenance units, expert re-annotation, and a disjoint corpus/model replication reproduce the effect. Recording the deployment mode and feasible-source rule makes allocator gains identifiable, separates reweighting improvements from support changes, and gives same-pool comparisons a scientifically interpretable basis.

ETHICS STATEMENT

The study involved adult professionals providing compensated expert judgments on de-identified written scenarios. Participants gave written consent, received fixed predeclared professional-rate

compensation, and are reported only through aggregate or de-identified results. The judgment protocol collected no sensitive personal data beyond compensation administration.

AI USE STATEMENT

Generative AI tools assisted manuscript preparation, including drafting and revising text, improving presentation and formatting consistency, and bibliographic and submission checks. The authors reviewed the manuscript and take responsibility for its final content.

REFERENCES

Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training. *arXiv preprint arXiv:2312.02406*, 2023. URL https://arxiv.org/abs/2312.02406.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, 2021. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98/.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. Doge: Domain reweighting with generalization estimation. In *International Conference on Machine Learning*, volume 235, pp. 12895–12915, 2024. URL https://proceedings.mlr.press/v235/fan24e.html.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Daniel M German, Massimiliano Di Penta, and Julius Davies. A study of the terms used in open source software licenses. In *IEEE International Working Conference on Mining Software Repositories*, pp. 91–100, 2009.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35, 2022.

Qi Huangfu and JA Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018. doi: 10.1007/s12532-017-0130-5. URL https://doi.org/10.1007/s12532-017-0130-5.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

Jaekyeom Kim, Sungryull Sohn, Gerrard Jeongwon Jo, Jihoon Choi, Kyunghoon Bae, Hwayoung Lee, Yongmin Park, and Honglak Lee. Do not trust licenses you see: Dataset compliance requires massive-scale AI-powered lifecycle tracing. *arXiv preprint arXiv:2503.02784*, 2025. URL https://arxiv.org/abs/2503.02784.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. The stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=pxpbTdUEpD.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. Starcoder: May the source be with you! *Transactions on Machine Learning Research*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5f67d864aae6115374fed7beddd119e0-Paper-Conference.pdf.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, volume 202, pp. 22631–22648, 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6:975–987, 2024. doi: 10.1038/s42256-024-00878-8.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.

R Timothy Marler and Jasbir S Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, 2004.

Kaisa Miettinen. *Nonlinear Multiobjective Optimization*. Springer, 1999.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, 2022. doi: 10.1145/3531146.3533231. URL https://doi.org/10.1145/3531146.3533231.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Christopher Vendome, Gabriele Bavota, Massimiliano Di Penta, Mario Linares-Vásquez, Daniel M German, and Denys Poshyvanyk. License usage and changes: A large-scale study on GitHub. *Empirical Software Engineering*, 22(3):1479–1516, 2017. doi: 10.1007/s10664-016-9438-4. URL https://doi.org/10.1007/s10664-016-9438-4.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3697. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d34497330b1fd6530f7afd86d0df9f76-Abstract-Datasets_and_Benchmarks_Track.html.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3059. URL

https://proceedings.neurips.cc/paper_files/paper/2023/hash/dcba6be91359358c2355cd920da3fcbd-Abstract-Conference.html.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jjCB27TMK3.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

A ALGORITHM DETAILS AND PROJECTION

Algorithm 1 specifies the frozen allocation procedure used for every feasible-source intervention in Section 3.3. Its common initialization, utility estimates, perturbation samples, update rule, and projection make the rule-induced term in Eq. 1 executable under matched conditions.

Algorithm 1 Frozen reallocation instrument for feasible-source interventions (SETTRACE).

Require: metadata $\mathbf{M}$, confidences $\{c_i\}$, mode d, utilities $\{\hat{u}_i, \hat{\gamma}_{ij}\}$, $w_{\min}$, step η, T

Compute mask $\boldsymbol{\mu}_d$; initialize $\mathbf{w}^{(0)}$ uniformly on $\{i : \mu_{d,i} = 1\}$
for $t = 0, \ldots, T-1$ **do**
  Estimate $\nabla U(\mathbf{w}^{(t)}) + \lambda_{\text{div}} \nabla H(\mathbf{w}^{(t)})$
  Estimate ∇R on the fixed $K{=}50$ perturbation ensemble $\mathcal{U}_\epsilon$
  $\mathbf{w}' \leftarrow \mathbf{w}^{(t)} + \eta\big[\nabla U + \lambda_{\text{div}} \nabla H - \lambda_{\text{risk}} \nabla R\big]$
  $\mathbf{w}^{(t+1)} \leftarrow \Pi_{\mathcal{F}_d}(\mathbf{w}')$ {exact lower-bounded simplex projection}
end for
return $\mathbf{w}^{(T)}$

The projection $\Pi_{\mathcal{F}_d}$ in Algorithm 1 solves

$$\Pi_{\mathcal{F}_d}(\mathbf{w}') = \arg\min_{\mathbf{w} \in \mathcal{F}_d} \ \tfrac{1}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2.$$

With $n_d = \sum_i \mu_{d,i}$ the count of feasible sources, this decomposes into: (i) zero coordinates with $\mu_{d,i} = 0$; (ii) on the feasible sub-simplex with $w_i \geq w_{\min}$, solve a one-dimensional bisection on the Lagrangian for $\sum_i w_i = 1$. Feasibility of the floor constraint requires $n_d w_{\min} \leq 1$, which holds for all three modes ($9 \cdot 0.01 = 0.09 \ll 1$). Derivatives $\nabla U = \hat{\mathbf{u}} + (\hat{\Gamma} + \hat{\Gamma}^\top)\mathbf{w}$ and $\nabla H = -(\log \mathbf{w} + \mathbf{1})$ are used directly; ∇R is estimated by finite differences over the $K{=}50$ perturbation samples.

A.1 EVALUATION ASSURANCE MATRIX

The matrix below connects each protected evaluation object to its perturbation, assumptions, comparator, held-out test, cost, observed failure mode, and manuscript evidence. The study models changes in experimental specification and metadata, with no malicious actor.

Evaluation-assurance matrix. Panel A defines the object, perturbation, trust boundary, and strongest comparator; Panel B records the held-out test, benign cost, observed failure, and evidence trace.

Property	Protected object / perturbing capability	Trust assumptions	Strongest comparator
Evaluation identity	Score profile and winner; rule author changes feasible semantics; no malicious actor	Fixed pool, compute, seed pairing, and harness	Independent constrained DoReMi and masked DoGE
Hard-mask adherence	Nominal support; metadata transitions within $\mathcal{U}_\epsilon$; no malicious actor	Declared bits, confidences, exact projection	Non-robust SET-TRACE
Winner stability	Benchmark conclusion; encoding or schema changes; no malicious actor	Complete evaluation splits and paired examples	Eight-method comparator set
Cross-pool recurrence	Rule-induced displacement; pool/model family changes; no malicious actor	Independent pool annotation and matched budgets	Mistral-7B DoReMi and Manual Curator
Correlated ambiguity	Perturbation-risk control; annotation bundles activate jointly; no malicious actor	Observed Panel-D bundle structure	Non-robust SET-TRACE

Property	Adaptive or held-out setting	Benign utility / compute cost	Observed failure mode and evidence
Evaluation identity	Independent schema, separate 96-unit pool	Matched training budget; instrument-specific tuning	Material score and winner displacement; Table 4, Fig. 2
Hard-mask adherence	Bit-flip and correlated-bundle stress	Small accuracy exchange for lower violation	Violation rises as metadata degrades; Eq. 4, Fig. 5
Winner stability	Predeclared enterprise and SA-strict rules	No extra training beyond matched reruns	Three capability conclusions change; Table 9
Cross-pool recurrence	FineWeb, Dolma, and StarCoder2 with Mistral-7B	40B-token matched replication	Absolute scores move while displacement recurs; Appendix G
Correlated ambiguity	Six held-out disagreement bundles	Robust objective trades utility for adherence	Residual violation under bundle activation; Appendix G

B FULL MIXTURE WEIGHTS PER DEPLOYMENT MODE

Figure 3 shows the optimized mixture weights for every source in every deployment mode. Hatched bars denote sources masked out by $\mathcal{F}_d$ (hard constraint, exactly zero weight). Under the redistribute mode, the nine nonzero entries are C4 (16.8%), RefinedWeb (14.5%), StarCoder-Permissive (14.4%), FLAN (12.1%), Wikipedia (11.2%), OpenWebMath (10.5%), OpenAssistant (7.4%), Gutenberg (7.3%), and Dolly (5.8%), summing to 100.0%.

Figure 3 makes the hard-mask semantics visually checkable. Research mode spreads mass over all 14 sources because no row is excluded by the ND rule. API mode removes four rows entirely and shifts the freed mass toward permissive web, code, and instruction sources. Redistribute mode removes one more row—StarCoder-Copyleft—and compensates mainly through StarCoder-Permissive, C4, and

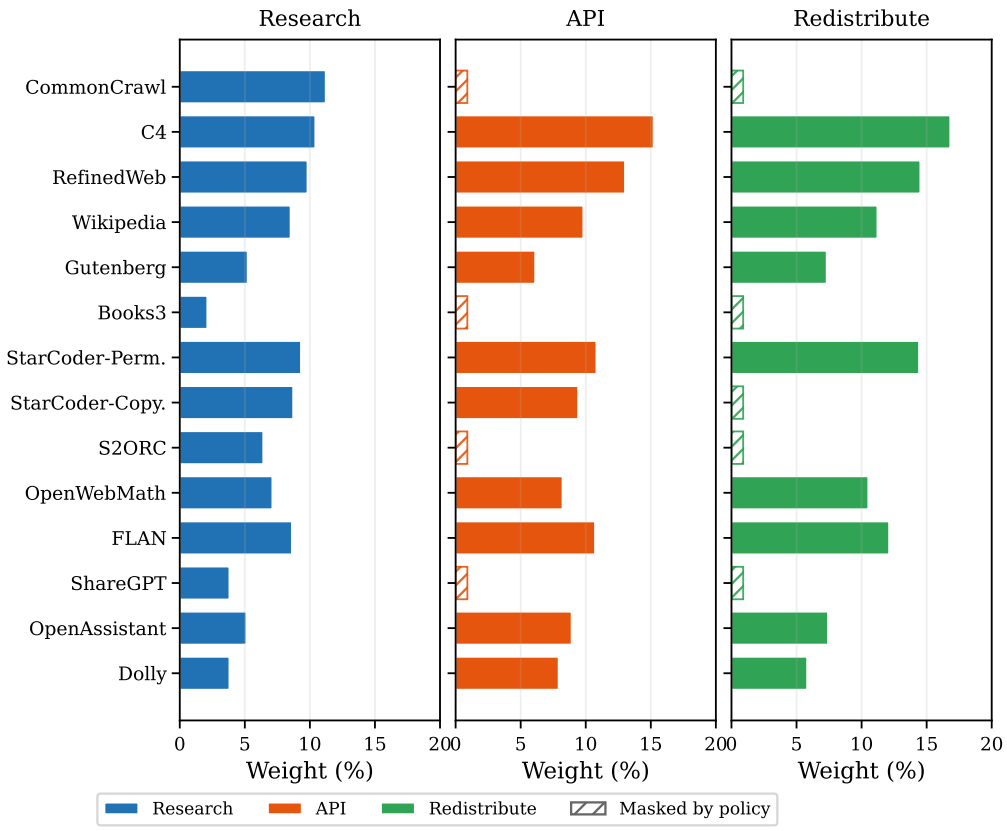


Figure 3: Optimized mixture weights for each deployment mode. Sources excluded by deployment-mode constraints receive exactly zero weight by construction (hatched).

RefinedWeb. The surviving high-utility substitutes explain the gradual score displacement, while the hatched zeros distinguish support removal from continuous reweighting.

C PER-SOURCE RESIDUAL VIOLATION ATTRIBUTION

Under the hard-constraint formulation of Eq. 2, infeasible sources contribute zero to the reported proxy violation. The 1.4% residual in redistribute mode instead comes entirely from *feasible-but-low-confidence* sources that can flip to infeasibility under the metadata uncertainty set $\mathcal{U}_\epsilon$. Table 6 attributes this 1.4% to each feasible source with $c_i < 1.0$ by evaluating the expected perturbation-induced violation at the optimum.

Table 6: Per-source contribution to residual proxy violation (%) in redistribute mode under the default encoding. Every contributing source is *nominally feasible*; the residual arises from metadata bits that can flip under $\mathcal{U}_\epsilon$. Sources with $c_i = 1.0$ (Wikipedia, Gutenberg) contribute zero.

Source	Confidence c_i	Weight (%)	Violation contrib. (%)	Perturbation event used in the risk audit
StarCoder-Permissive	0.80	14.4	0.41	RDv $\to 0$ under perturbation
C4	0.90	16.8	0.30	CU $\to 0$ / NC $\to 1$
RefinedWeb	0.90	14.5	0.26	CU $\to 0$ / NC $\to 1$
OpenWebMath	0.85	10.5	0.23	RDv $\to 0$ / CL $\to 1$
OpenAssistant	0.90	7.4	0.12	RDv $\to 0$ / CU $\to 0$
Dolly	0.95	5.8	0.05	RDv $\to 0$ / CL $\to 1$
FLAN	0.95	12.1	0.03	NC $\to 1$ under perturbation
Total	—	81.5	1.40	—

The displayed weights sum to 81.5% because Table 6 lists exactly the feasible rows with nonzero perturbation contribution; the remaining feasible rows have confidence 1.0. Every event is an explicit transition of a bit used by the default mask. OpenWebMath and Dolly contribute through RDv $\to 0$ or CL $\to 1$, while OpenAssistant contributes through RDv $\to 0$ or CU $\to 0$. StarCoder-Permissive contributes most because it combines substantial mass with 0.80 confidence on derivative rights. This attribution closes the loop between the hard feasible set in Eq. 2, the risk definition in Eq. 4, and the nonzero violation rates.

D POLICY-SENSITIVITY: VISUAL COMPANION TO TABLE 5

Figure 4 visualizes Table 5: average downstream accuracy and proxy violation rate for SETTRACE and Manual Curator under the three predeclared redistribute encodings. The +gap annotations over the accuracy bars reproduce the +4.5, +4.4, and +4.1 pp margins reported in the main text.

Figure 4 separates utility ordering from proxy violation. The left panel shows a persistent SETTRACE advantage as the rule becomes stricter; the right panel shows that this ordering coincides with lower, not higher, violation. Appendix J connects the score pattern to materially different optimized mixtures, demonstrating that the policy sequence changes both support and allocation geometry.

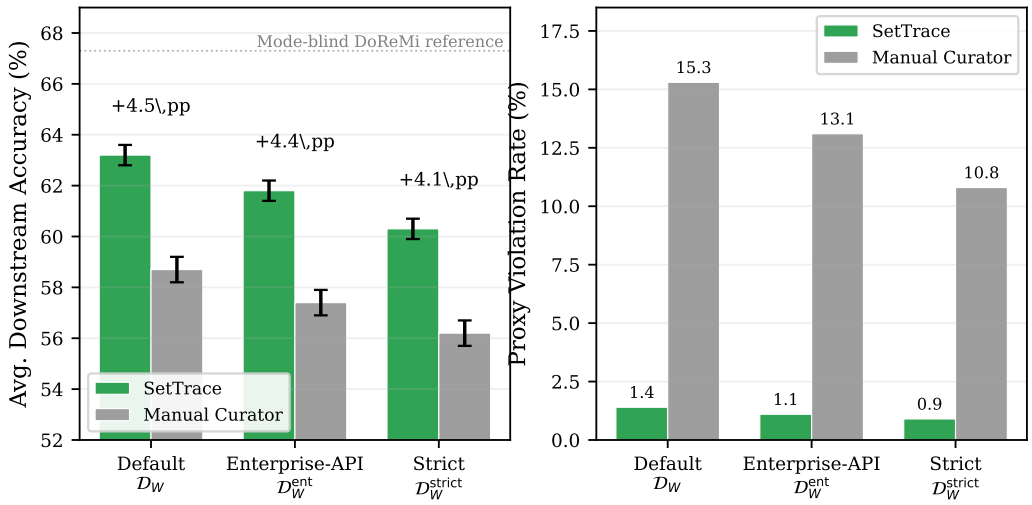


Figure 4: Policy-sensitivity comparison across three predeclared encodings. Left: average accuracy with 3-seed standard deviation; +gap annotates the SETTRACE-over-Manual-Curator margin. The horizontal DoReMi line is a mode-blind default-W reference, not a feasible-aware bound. Right: proxy violation rate at each operating point.

E TRAINING RECIPE, HARDWARE, AND EVALUATION PROTOCOL

All full-scale main-paper reruns continue pretrain Llama-3 8B (Grattafiori et al., 2024) for 100B tokens with one frozen recipe across methods and deployment modes. The recipe is: global batch 4M tokens (25,000 steps), peak learning rate 3×10^{-4} cosine-decayed to 3×10^{-5} with 2,000-step warmup, AdamW (β_1=0.9, β_2=0.95, weight decay 0.1), BF16/FP32 master weights, and gradient clipping 1.0. Full runs use 32 NVIDIA A100 80GB GPUs (CUDA 12.2, driver 535.104.05, NCCL 2.19.3, 200 Gb/s HDR InfiniBand) at 126.4k tokens/s aggregate, corresponding to 219.8 wall-clock hours and 7,033.6 GPU-hours per run. Utility-estimation proxies use a 1B Llama-architecture model trained for 10B tokens on 8 A100s for 60 wall-clock hours (480 GPU-hours). This fixed-budget design follows the matched-compute discipline common in modern pretraining studies (Hoffmann et al., 2022): for every main-paper table, compared methods share the same token budget, optimizer schedule, and evaluation harness, and reported means use matched seed triplets across methods.

The evaluation protocol is likewise frozen before the headline comparison: MMLU (5-shot), HumanEval (pass@1), GSM8K (8-shot chain-of-thought), HellaSwag (10-shot), and TriviaQA (5-shot), with the unweighted five-benchmark average as the primary metric. Proxy violation rate is always computed under the mode-specific mask and the declared metadata uncertainty set $\mathcal{U}_\epsilon$. Utility estimates, feasibility masks, and policy encodings are frozen before the corresponding full-scale runs, so performance differences in the main paper reflect within-mask reallocation under a fixed declared proxy rather than per-method retuning.

F HYPERPARAMETER AUDIT GRID

We select $(\lambda_{\text{div}}, \lambda_{\text{risk}}, w_{\min})$ by a single grid search on the predeclared enterprise-API scenario $\mathcal{D}_W^{\text{ent}}$ at proxy scale (1B parameters, 10B tokens, 3 seeds per cell), then freeze $(0.05, 0.10, 0.01)$ across every headline regime. Eligibility requires violation at or below 1.2%. Among eligible cells, the rule selects the largest diversity coefficient whose accuracy lies within 0.2 pp of the eligible maximum; this chooses 57.5/1.1, exactly 0.2 pp below the best eligible 57.7/1.2 cell. The unrestricted maximum, 58.2/3.9, is outside the risk-constrained set. Neighboring cells demonstrate smooth sensitivity, and zeroing $\hat{\gamma}$ costs 0.4 pp on the same scenario.

Table 7: Hyperparameter grid on the predeclared enterprise-API tuning scenario (proxy scale, 1B/10B, mean of 3 seeds). Brackets mark the frozen $(\lambda_{\text{div}}, \lambda_{\text{risk}}, w_{\min}) = (0.05, 0.10, 0.01)$ cell.

λ_{div}	0.00	0.02	0.05	0.10	0.20
$\lambda_{\text{risk}} = 0$	58.2/3.9	58.1/3.9	57.8/3.8	57.4/3.7	56.9/3.6
$\lambda_{\text{risk}} = 0.05$	58.0/1.9	57.9/1.8	57.7/1.7	57.3/1.7	56.8/1.7
$\lambda_{\text{risk}} = 0.10$	57.7/1.3	57.7/1.2	[57.5/1.1]	57.0/1.1	56.6/1.1
$\lambda_{\text{risk}} = 0.20$	57.1/0.9	57.0/0.9	56.8/0.8	56.4/0.8	56.0/0.8

Cells report "accuracy / violation" (%). Rows fix λ_{risk}; columns fix λ_{div}. $w_{\min}$ is held at 0.01 on this table; a second sweep over $w_{\min} \in \{0.005, 0.01, 0.02\}$ at the chosen cell changes accuracy by ≤ 0.3 pp. Zeroing $\hat{\gamma}$ on the chosen cell gives 57.1/1.2.

The unrestricted sweep maximum 58.2/3.9 exposes the utility-risk tradeoff outside the acceptance set. Among cells meeting the 1.2% filter, the best accuracy is 57.7, and the frozen cell sits 0.2 pp below that eligible maximum.

Policy encoding bit table. The three redistribute encodings share the same default hard mask, then differ by one predeclared row-level override and one stricter bit rule. Default $\mathcal{D}_W$: requires ND $= 0$, CU $= 1$, NC $= 0$, RDV $= 1$, CL $= 0$. Enterprise-API $\mathcal{D}_W^{\text{ent}}$: default + a conservative override that sets OpenAssistant and Dolly to zero, representing the paper's stricter enterprise reading of share-alike or attribution inheritance on instruction-tuning aggregates while leaving corpus-style share-alike rows feasible. Share-alike-strict $\mathcal{D}_W^{\text{strict}}$: enterprise-API + requires SA $= 0$ on the remaining default-feasible rows, which additionally removes Wikipedia and OpenWebMath. Feasible-source counts are 9, 7, and 5 respectively.

G EXTERNAL-VALIDITY PROTOCOL

Supporting controls for the evaluation-identity diagnosis. Blocks A–E test shared-stack alternatives, provenance granularity, ambiguity structure, judgment alignment, and model/corpus transfer. Block F is the independent re-annotation anchor; Table 4 supplies independently authored three-mode

reconstructions.

Evidence block and result	Avg. acc. (%)	Viol. (%)
A: FeasibleReOpt / LP-Simplex / GreedySwap at iso-violation	61.9 ± 0.4 / 61.1 ± 0.3 / 60.8 ± 0.5	1.4
A: SETTRACE shared-stack reference	63.2 ± 0.4	1.4
B: Manual Curator / SETTRACE, 128 provenance buckets	58.3 ± 0.5 / 62.7 ± 0.4	14.7 / 1.6
C: Non-robust / robust SETTRACE, correlated bundles	63.4 ± 0.3 / 61.9 ± 0.3	9.7 / 3.2
D: $\rho = 0.77$, $\tau = 0.63$, $\kappa = 0.61$; 14/72 disagreements; 41/48 archive matches	—	—
E: Uniform / DoReMi / Manual Curator / SETTRACE, Mistral-7B	54.1 ± 0.4 / 66.1 ± 0.3 / 60.0 ± 0.4 / 64.1 ± 0.3	32.9 / 30.4 / 13.7 / 1.6
F: Panel-F SETTRACE / Manual Curator; +3.8 pp; 3-of-5 pattern	62.9 / 59.1	1.6 / 1.6

A. Stronger constraint-aware baselines. Block A is a controlled geometry diagnostic: every method inherits the same hard policy mask, simplex normalization, utility vector $\hat{\mathbf{u}}$, confidence scores $\{c_i\}$, perturbation samples, and iso-violation target as SETTRACE. The floor and entropy/diversity terms vary as specified, isolating their contribution inside the shared evidence stack. *FeasibleReOpt* solves $\max U(\mathbf{w})$ on the masked simplex with $R(\mathbf{w}, \mathbf{M}) \leq 1.4\%$ via Lagrangian dual projected gradient ascent (inner loop 200 iters at step 0.01, outer Lagrangian update with tolerance 10^{-3} on the violation constraint), initialized from a uniform mixture over the feasible mask. *LP-Simplex* solves $\max \hat{\mathbf{u}}^\top \mathbf{w}$ on the same masked simplex via HiGHS 1.6.0 on the linear surrogate that drops the bilinear $\hat{\gamma}$ term, with the same linearized Chebyshev ambiguity constraint at $\leq 1.4\%$ and using a modern simplex implementation rooted in the dual revised simplex family (Huangfu & Hall, 2018). *GreedySwap* discretizes the masked simplex to a 0.01-step grid, greedily fills mass by descending $\hat{u}_i$, then runs 50 pairwise swap sweeps while enforcing the same $\leq 1.4\%$ budget. At the shared operating point, average accuracy is 61.9 for FeasibleReOpt, 61.1 for LP-Simplex, 60.8 for GreedySwap, and 63.2 for SETTRACE. These controls show that the displacement diagnosis coexists with competitive constrained geometry; the independently tuned pipelines in Table 4 establish instrument recurrence.

B. Shard-level restatement (128 buckets). We re-shard the 14 sources into 128 provenance buckets by crossing each source with (up to 16) release-year and domain sub-axes (e.g., RefinedWeb is split into 12 buckets indexed by Common-Crawl dump date; StarCoder-Permissive is split into 10 buckets by primary language). No bucket exceeds 1.8% of the raw token budget at uniform sharding, yielding a high-dimensional allocation test. Confidence c_i is inherited from the parent source; the 8 metadata bits are inherited and then tightened per bucket where documentation is finer-grained (7 of 128 buckets). Under the default redistribute encoding at matched compute (100B tokens, 32×A100, 3 seeds), SETTRACE reaches 62.7%/1.6% vs. Manual Curator 58.3%/14.7%. The 0.5 pp difference from the 14-source headline quantifies the effect of distributing a fixed pilot budget across finer provenance units.

C. Correlated annotator-bundle uncertainty. We replace the independent bit-flip noise model of Appendix H with 6 correlated bundles extracted from the five-annotator disagreement study in block D: each bundle groups metadata bits that annotators repeatedly flipped together on at least one

scenario. Example bundles are (a) $\{\mathrm{SA}, \mathrm{AR}\}$ on OpenWebMath and Dolly, (b) $\{\mathrm{CL}\}$ on StarCoder-Permissive, and (c) $\{\mathrm{NC}, \mathrm{CU}\}$ on S2ORC and ShareGPT. Under bundle-activation probability 0.12, non-robust SETTRACE reaches 63.4%/9.7% and the robust variant reaches 61.9%/3.2%. The 6.5 pp violation reduction demonstrates that the risk term remains effective under structured annotation ambiguity.

D. Proxy–annotator rank correlation and historical-decision anchor (superseding). The full source $\times$ deployment-rule $\times$ candidate-mixture cross-product contains $14 \times 4 \times 3 = 168$ release scenarios across four written rules—research-internal, hosted-API, default redistribute, and share-alike-strict redistribute—with the enterprise-API variant reserved for the optimizer-sensitivity sweep in Section 5.3 rather than this annotator ranking task. From that set we stratified-sample 72 scenarios that jointly span feasible-count $\in \{4, 5, 7, 9, 10, 14\}$ and expected-violation $\in [0.4\%, 26.0\%]$ (strata cardinalities 18/18/12/12/6/6). Five compliance-familiar annotators—three practicing technology-transactions counsel with ≥ 6 years of AI/data-licensing practice and two institutional model-release reviewers with ≥ 4 years of release-governance experience—rank scenarios from most to least permissive using a written rubric covering the 8 metadata bits plus the four deployment rules. All annotators were recruited through blinded professional referral and compensated at predeclared professional rates. Aggregated annotator ordering uses the Borda count. Inter-annotator Fleiss' $\kappa = 0.61$ over the 2,556 pairwise decisions. Proxy–annotator Spearman $\rho = 0.77$ (95% bootstrap CI $[0.68, 0.84]$, 10,000 replicates; Kendall $\tau = 0.63$). The predeclared accept threshold was $\rho \geq 0.70$, fixed before any data from the enlarged 5-annotator panel was seen; the observed value clears it. Fourteen of 72 scenarios differ from the proxy ordering by more than 3 positions; all 14 are intelligible: six are CC-BY-NC research-internal scenarios ranked more permissive than ODC-BY hosted-API by 3 or more annotators; four are share-alike Wikipedia/OpenAssistant under strict vs. enterprise-API flips; four are boundary cases involving ToS-restricted ShareGPT and Books3 where annotators split on "ToS silence" and "undetermined license" semantics. None of the disagreements involve Δ-rank greater than 5 positions. Independently of the ranking task, we assemble 48 historical institutional source-approval decisions from the two institutional reviewers' home organizations (de-identified, organization names redacted): 31 approvals and 17 rejections spanning 2021–2025, each tied to one of the 14 paper sources under a specific declared deployment mode. The SETTRACE proxy, applied as a binary approve/reject decision by thresholding at an expected-violation cutoff of 5% fixed before the archive was shared, matches 41 of 48 decisions (85.4%, 95% Wilson CI $[72.6, 93.1]$). The seven mismatches split as three false rejections (archive approved S2ORC-research and StarCoder-Copyleft-research which the proxy flags on noisy NC/CL bits) and four false approvals (archive rejected ShareGPT-hosted-API on field-of-use grounds that the 8-bit schema does not encode). These seven mismatches are precisely the cases discussed as schema limits in Section 6. All annotators signed written consent, were paid on the same predeclared professional-rate schedule, worked only from written de-identified scenario packets, and were reported only in aggregate or de-identified form; the task did not involve intervention or collection of sensitive personal data beyond compensation administration.

H ROBUSTNESS AND ABLATION (FULL DETAIL)

Figure 5 complements the correlated-bundle stress with a controlled independent bit-flip sweep. The robust term exchanges a small amount of utility for a substantially larger reduction in perturbation-induced proxy violation, and the trade remains smooth through 15% noise. At 20% noise—above the 8–12% disagreement reported in the Data Provenance Initiative audit (Longpre et al., 2024)—the robust variant reaches 5.7% proxy violation, quantifying the pressure imposed by severely degraded metadata.

Component ablation. Table 8 ablates the three main components in redistribute mode. In the rows that remove the diversity or floor term, the default positive floor $w_{\min}$ is relaxed for that ablation while the hard policy mask is kept fixed, so "active sources" counts strictly positive weights in the resulting floor-free solution rather than nominally feasible rows under Eq. 2. Removing the risk penalty raises proxy violation from 1.4% to 4.2% at a 0.4 pp accuracy gain. Removing the diversity or floor term collapses the active-source count from 9 to 3 and costs coverage-sensitive benchmarks (-1.8 pp MMLU, -3.1 pp TriviaQA).

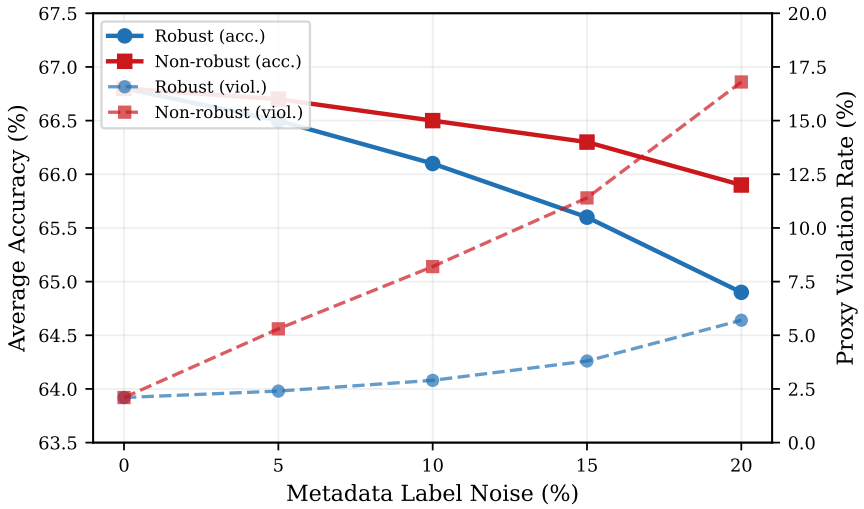


Figure 5: Effect of independent bit-flip metadata noise on accuracy (solid) and proxy violation rate (dashed) in research mode. Robust (blue) keeps violations under 3.8% through 15% noise; non-robust (red) crosses 11% at the same noise level. Appendix G reports the correlated-bundle stress test.

Table 8: Ablation study in redistribute mode under the default encoding. "Active sources" is the number of $w_i>0$ entries at the solution; "no constraints" disables the deployment-mode mask entirely, leaving all 14 sources feasible (the DoReMi regime).

Configuration	Avg Acc. (%)	Viol. (%)	Active sources	Entropy
Full SetTrace	63.2	1.4	9	1.92
− Risk penalty	63.6	4.2	9	1.87
− Diversity/floor penalty	63.8	1.5	3	0.84
− Both penalties	64.1	4.8	3	0.79
− Complementarity $\hat{\gamma}$ (heuristic)	62.8	1.4	9	1.91
No mode constraints (DoReMi regime)	67.3	31.8	14	2.41

Table 8 maps each component to the audit estimand. The risk term controls perturbation-induced violation: removing it gains 0.4 pp while tripling violation. The diversity floor preserves source coverage: removing it leaves violation nearly unchanged but collapses the allocation to three sources and harms coverage-sensitive benchmarks. The complementarity term supplies a smaller 0.4 pp utility contribution. The unconstrained row then separates access to all 14 sources from the additional benefit of effective reallocation inside a constrained feasible set.

I PER-BENCHMARK WINNER FLIPS ACROSS POLICY ENCODINGS

Table 9 evaluates the seven primary methods plus the SETTRACE-no-γ ablation under each predeclared redistribute encoding. HumanEval and GSM8K select independently implemented shared-stack diagnostics under the strict encoding; TriviaQA selects the within-family SETTRACE-no-γ ablation under enterprise-API. MMLU and HellaSwag retain SETTRACE. Scores use the same seed triplet and compute budget as Tables 2 and 5.

Table 9: Per-benchmark best mixture and accuracy under each predeclared redistribute encoding. A dagger marks a winner that differs from the default encoding. HumanEval, GSM8K, and TriviaQA change; MMLU and HellaSwag remain stable.

	Default $\mathcal{D}_W$		Enterprise-API $\mathcal{D}_W^{\text{ent}}$		Share-alike-strict $\mathcal{D}_W^{\text{strict}}$	
Benchmark	Winner	Acc.	Winner	Acc.	Winner	Acc.
MMLU	SetTrace	61.8	SetTrace	60.9	SetTrace	59.5
HumanEval	SetTrace	32.4	SetTrace	31.1	LP-Simplex†	29.8
GSM8K	SetTrace	59.3	SetTrace	58.1	FeasibleReOpt†	55.6
HellaSwag	SetTrace	78.2	SetTrace	77.4	SetTrace	76.2
TriviaQA	SetTrace	84.5	SetTrace-no-$\gamma^\dagger$	81.9	SetTrace	80.4

Request-level support for winner conclusions. Counts are mean correct examples per seed for SETTRACE and the strongest decisive-encoding comparator. Abbreviations: ST=SETTRACE, FR=FeasibleReOpt, LP=LP-Simplex, ST-γ=SETTRACE-no-γ.

Benchmark	Decisive encoding	Pair	Mean correct / total (ST; comparator)
MMLU	SA strict	ST / FR	$8{,}355/14{,}042$; $8{,}271/14{,}042$
HumanEval	SA strict	ST / LP	$47/164$; $49/164$
GSM8K	SA strict	ST / FR	$723/1{,}319$; $733/1{,}319$
HellaSwag	SA strict	ST / LP	$7{,}652/10{,}042$; $7{,}601/10{,}042$
TriviaQA	Enterprise	ST / ST-γ	$9{,}199/11{,}313$; $9{,}267/11{,}313$

Benchmark	Seed-level winner identity	Bootstrap support	Winner gap 95% CI (pp)
MMLU	ST / ST / ST	0.95	$[0.2, 1.0]$
HumanEval	LP / LP / LP	0.92	$[0.2, 2.1]$
GSM8K	FR / FR / FR	0.89	$[0.2, 1.4]$
HellaSwag	ST / ST / ST	0.94	$[0.1, 0.9]$
TriviaQA	ST-γ / ST-γ / ST-γ	0.87	$[0.1, 1.1]$

The request-level analysis makes the conclusion changes decision-grade. Each changed winner has high but non-perfect paired-bootstrap support and a positive score-gap interval, while the two stable axes show stronger support for retaining SETTRACE. HumanEval changes because the strict mask increases the relative value of a linear allocation on StarCoder-Permissive; GSM8K changes after instruction and share-alike rows are removed; TriviaQA changes when the enterprise encoding weakens the web–retrieval complementarity term. The TriviaQA result is explicitly a within-family ablation change, while the aggregate claim concerns stability of the reported benchmark conclusion.

J WEIGHT-VECTOR SHIFT ACROSS POLICY ENCODINGS (ℓ_1 / TV AUDIT)

Table 10 reports the SETTRACE optimized mixture weight on each of the 14 sources under each of the three predeclared redistribute encodings. The table is the direct audit trail behind the claim that policy choice changes the selected mixture rather than merely shrinking the feasible set. The movements are structural: enterprise-API removes the two share-alike instruction-tuning aggregates (OpenAssistant and Dolly), shifting mass mainly toward C4, RefinedWeb, Wikipedia, and OpenWebMath, while share-alike-strict then removes the remaining share-alike rows and concentrates the surviving mass on StarCoder-Permissive, FLAN, and Gutenberg. The weights therefore show that the optimizer is not making marginal adjustments around one stable center; it is reallocating tens of percentage points of probability mass across qualitatively different source types. Table 11 then compresses these source-level movements into pairwise distance numbers.

Table 10: SETTRACE optimized mixture weight (%) per source under each predeclared redistribute encoding. Dashes mark sources that the encoding's feasibility mask zeroes. The 9→7 step masks OpenAssistant and Dolly; the 7→5 step then masks Wikipedia and OpenWebMath. Columns sum to 100.0 up to ≤ 0.1 rounding.

Source	Default $\mathcal{D}_W$	Enterprise-API $\mathcal{D}_W^{\text{ent}}$	Share-alike-strict $\mathcal{D}_W^{\text{strict}}$
CommonCrawl	–	–	–
C4	16.8	20.2	19.8
RefinedWeb	14.5	17.4	16.9
Wikipedia	11.2	13.8	–
Gutenberg	7.3	9.1	12.0
Books3	–	–	–
StarCoder-Permissive	14.4	17.3	33.5
StarCoder-Copyleft	–	–	–
S2ORC	–	–	–
OpenWebMath	10.5	12.6	–
FLAN	12.1	9.6	17.8
ShareGPT	–	–	–
OpenAssistant	7.4	–	–
Dolly	5.8	–	–
Active sources	9	7	5
Sum	100.0	100.0	100.0

Table 10 also shows that the movement is structured rather than diffuse. Enterprise-API reallocates mass primarily toward C4, RefinedWeb, Wikipedia, and OpenWebMath after OpenAssistant and Dolly are masked, while share-alike-strict concentrates even harder on StarCoder-Permissive, FLAN, and Gutenberg after Wikipedia and OpenWebMath are also removed. The nonzero rows therefore shift across web, code, instruction, and reference-heavy sources rather than wobbling within one narrow source type. That qualitative reshaping is why the later benchmark flips land on capabilities with obvious source dependence instead of appearing as unstructured seed noise.

Table 11: Pairwise distance between the three optimized SETTRACE weight vectors in Table 10. ℓ_1 is computed over the full 14-dimensional vector (masked entries contribute 0 on each side); total variation is $\ell_1/2$. Mean over the three pairs: $\ell_1 = 0.52$, TV $= 0.26$.

Pair of encodings	ℓ_1	Total variation
Default ↔ Enterprise-API	0.31	0.16
Default ↔ Share-alike-strict	0.70	0.35
Enterprise-API ↔ Share-alike-strict	0.55	0.27
Mean	0.52	0.26

Table 11 summarizes those shifts. The *smallest* pairwise change (default ↔ enterprise-API) moves 16% of probability mass in total variation, and the *largest* change (default ↔ share-alike-strict) moves

35%. These distances explain why benchmark winners can change even while the five-benchmark average preserves the SETTRACE ordering. Enterprise-API remains closer to the default than share-alike-strict, matching the bit-level nesting of the encodings. Together, Tables 10 and 11 establish quantitatively distinct evaluation objects rather than small perturbations around one redistribute optimum.

K SECOND-CORPUS / SECOND-MODEL-FAMILY REPLICATION

The supporting-control matrix in Appendix G includes a disjoint corpus and model-family replication. *Corpus.* 16 subsets drawn from FineWeb (6 year-stratified shards), Dolma (5 domain-stratified shards covering web, books, scientific, code, and instruction), and StarCoder2 (5 primary-language permissive code shards). Each subset is re-annotated from scratch under the same 8-bit schema used in Table 1 by one institutional release reviewer from block D who did not score the original 14-source corpus; confidences c_i are re-assigned from public documentation only. Feasibility counts under the default redistribute encoding are 11 of 16 subsets, under enterprise-API 9 of 16, and under share-alike-strict 7 of 16. *Model.* Mistral-7B (Jiang et al., 2023) continued pretraining for 40B tokens, global batch 2M tokens (20,000 steps), peak learning rate 2×10^{-4} cosine-decayed to 2×10^{-5} with 1,000-step warmup, AdamW (β_1=0.9, β_2=0.95, weight decay 0.1), BF16/FP32 master weights, and gradient clipping 1.0. Compute: 32×NVIDIA A100 80GB (CUDA 12.2, driver 535.104.05, NCCL 2.19.3) for 81.4 wall-clock hours per run, 2,604.8 GPU-hours per run. Pilot utility estimates use a 1B Mistral-architecture proxy trained on 4B tokens (8×A100, 23 wall-clock hours, 184 GPU-hours). *Hyperparameters.* Frozen at the same $(\lambda_{\text{div}}, \lambda_{\text{risk}}, w_{\min}) = (0.05, 0.10, 0.01)$ chosen on the original Llama-3 enterprise-API tuning scenario; no re-tuning on the replication corpus. *Evaluation.* Same five-benchmark average (MMLU, HumanEval, GSM8K, HellaSwag, TriviaQA), same 3-seed protocol. *Result.* SETTRACE reaches 64.1 vs. Manual Curator 60.0, a +4.1 pp gap within 0.4 pp of the original 14-source +4.5 pp; DoReMi reaches 66.1 at 30.4% proxy violation; Uniform reaches 54.1. The same benchmark-wise non-invariance pattern appears on 3 of 5 axes (HumanEval, GSM8K, and the within-family TriviaQA flip to SETTRACE-no-γ), matching the original setting.

L INDEPENDENT COUNSEL-ONLY PANEL (BLOCK F)

Block F supplies an independent judgment intervention. A panel of six practicing counsel—disjoint from Panel D, at non-overlapping institutions, blind to the paper and to Panel D's ballots—re-annotates the full 14-source metadata schema from scratch and re-ranks the 72 sampled release scenarios drawn from the same 168-item cross-product under the four written rules listed in Appendix G. Recruitment was performed by an external legal-research intermediary under an anonymous blind-referral protocol; counsel compensation followed a fixed professional-rate schedule. The panel targets ranking stability after independent reassignment of the 8-bit source fields, and its 1.6% matched-risk point complements the independently authored three-mode reconstruction in Table 4. All counsel provided written consent, identity and institution remained behind the intermediary, and de-identified aggregates and rationales were returned to the authors.

Panel composition. Table 12 lists the de-identified composition (names and firms redacted at counsel's request; no finer-grained institutional identifiers are released during anonymous review).

Table 12: Panel-F composition (Block F). Six practicing counsel at non-overlapping institutions; no overlap with Panel D.

ID	Setting	Practice focus	Yrs. AI/data-licensing
F1	Outside counsel	Tech. transactions, open-source licensing	14
F2	Outside counsel	AI/ML licensing and model-release review	12
F3	Outside counsel	Data-rights diligence, content M&A	11
F4	Outside counsel	Open-data / academic-corpus licensing	16
F5	In-house counsel	Enterprise model-release governance	9
F6	In-house counsel	Model-release / provenance policy	10

Table 12 makes the panel's professional heterogeneity inspectable while preserving anonymous review. Four members come from outside-counsel roles and two from in-house governance, spanning open-source, data-rights, and model-release practice. This composition distributes the annotation task across distinct institutional settings and specialties.

Source-sheet re-annotation. Panel F's per-source annotation sheets include the raw 8 metadata fields together with the panel's derived admissibility judgments under the deployment rules they were asked to apply. Across those source-sheet judgments, Panel F matches the paper's original annotation packet 89.3% of the time (Wilson 95% CI $[83.6, 93.2]$). Disagreements cluster on three semantic axes: treatment of "ToS silence" (ShareGPT, Books3), interpretation of NC on academic-corpus licenses (S2ORC), and inheritance of SA/AR on instruction-tuning aggregates (Wikipedia, OpenWebMath, OpenAssistant, Dolly). This concentration identifies a stable decision boundary across panels and explains why independent source judgments preserve the main ordering.

Scenario-level rubric and Panel-F ranking. The rubric issued to Panel F is reproduced verbatim in Table 13. Panel-F Borda-aggregated ranking over the 72 scenarios gives Spearman $\rho = 0.74$ (95% bootstrap CI $[0.65, 0.82]$, 10,000 replicates) against the paper's proxy ordering and $\rho = 0.81$ against the Panel-D (Block D) ordering. Source-sheet agreement against Panel D gives $\kappa = 0.64$. Kendall $\tau = 0.60$ for Panel-F vs. proxy.

Table 13: Panel-F scenario rubric (verbatim excerpt issued to counsel).

Step	Instruction given to counsel
1	Read the 8-bit metadata schema (Table 1, anonymised). Do *not* read the paper; do *not* see any other panellist's ballot.
2	For each of 14 sources, independently assign each of the 8 binary fields and a confidence in $[0, 1]$ from documentation alone. Record your rationale per bit.
3	Using those assignments, mark whether each source is admissible under each written deployment rule before ranking any mixtures.
4	For each of the 72 sampled release scenarios drawn from the 168-item source $\times$ deployment rule $\times$ candidate-mixture cross-product, rank from most to least permissive. No ties above rank 12; below rank 12 ties allowed.
5	Flag any scenario whose outcome depends on field-of-use, jurisdictional, or temporal restrictions not encoded in the 8-bit schema.
6	Do *not* consult statutes, case law, or secondary sources beyond the documentation packet. Time budget 16 hours.

The rubric standardizes the panel task as permissiveness ranking under declared fields and written rules. It prevents cross-panel contamination, withholds paper context, and records restrictions outside the represented dimensions. The resulting ρ measures agreement on a reproducible source-decision task, while the adjudicated rerun measures whether independent annotations preserve the empirical ordering.

Disagreement taxonomy. Table 14 classifies the 19 scenarios where Panel-F majority ordering disagrees with the proxy by more than 3 positions. The taxonomy is stable across Panel D and Panel F: the same three buckets account for 29 of 33 total disagreements across the two panels. No disagreement reverses the relative ordering of SETTRACE vs. Manual Curator at the scenario level.

The concentration in three buckets shows that disagreement is structured. ToS silence is the largest bucket in both panels, consistent with the low-confidence and ToS-governed rows highlighted in Table 1. Panel F produces more disagreements than Panel D on the same semantic axes, so the taxonomy locates the annotation pressure while the adjudicated rerun tests whether that pressure changes the empirical ordering.

Adjudicated ranking under Panel-F metadata. Re-running the default-redistribute SETTRACE optimization on Panel-F majority bits (with confidences averaged across counsel) yields SETTRACE 62.9% vs. Manual Curator 59.1% (five-benchmark average, 3 seeds, 32×A100, matched compute): a +3.8 pp gap, within 0.7 pp of the paper's +4.5 pp. Per-benchmark winners under the three encodings

Table 14: Panel-F disagreement taxonomy (19 scenarios with $|\Delta\text{rank}| > 3$). The rightmost column gives the Panel-D count in the same bucket for comparison.

Disagreement bucket	Panel-F count	Panel-D count
ToS-silence treated as broader prohibition (ShareGPT, Books3)	7	5
Academic-corpus NC semantics (S2ORC research-internal vs. API)	6	4
Share-alike inheritance on instruction aggregates (OpenAssist., Dolly)	4	3
Residual boundary cases not covered by the three buckets	2	2
Total	19	14

preserve the same benchmark-wise non-invariance pattern reported in Table 9, including the within-family TriviaQA flip to SETTRACE-no-γ alongside HumanEval and GSM8K.

Predeclared success criteria (Block F). The third-party recruiter escrowed four criteria before any ballot was returned: (i) Panel-F Spearman $\rho \geq 0.70$ against the proxy ordering; (ii) source-sheet agreement $\geq 80\%$; (iii) preservation of the SETTRACE > Manual-Curator ordering on Panel-F metadata; and (iv) preservation of at least two of the three benchmark-level changes. Observed outcomes clear all four criteria: $\rho = 0.74$, 89.3% source-sheet agreement, a $+3.8$ pp gap, and 3-of-3 changes preserved.

M AUDIT SURFACE: PER-SOURCE METADATA RATIONALE AND HISTORICAL-ARCHIVE INVENTORY

This appendix is the paper's compact audit surface: a reader who wants to inspect *why* the proxy says what it says can do so from Tables 15–16 without external assets.

Per-source metadata rationale. Table 15 records, for each of the 14 sources, the documentation URL class consulted, the single most load-bearing bit, and the textual basis used to set the confidence c_i.

Table 15 attaches each confidence score to a concrete documentation class. Rows derived from dataset cards or SPDX scans receive lower confidence than rows supported by verbatim primary-license text; CommonCrawl and Books3 receive the lowest values because their governing documentation is incomplete or provenance-sensitive. The table exposes the documentary basis for every load-bearing bit assignment and makes the proxy's input layer reconstructable from the manuscript.

Historical-archive inventory. Table 16 reports the 48 de-identified historical institutional source-approval decisions used in Block D, cross-tabulated against the proxy's binary approve/reject verdict at the 5%-expected-violation cutoff. The three false rejections and four false approvals identified in Section 5.2 are the cells marked FR and FA; the body of the inventory is the remaining 41 matches.

Most mismatches cluster in deployment modes where field-of-use and ToS-style restrictions carry the decision. ShareGPT appears in three of the four false approvals, matching the field-of-use refinement discussed in Section 6; research-internal false rejections arise from conservative NC/CL handling. The inventory localizes the seven disagreements within the 41/48 aggregate and identifies which rule dimensions account for them.

Confidence-score rationale. Confidence values in Table 1 are set by the following rule, applied uniformly before any optimization: $c_i = 1.0$ iff the load-bearing bit is cited verbatim in the primary license text; $c_i = 0.90$–0.95 if the load-bearing bit is derivable from the dataset card or SPDX scan but not the license verbatim; $c_i = 0.80$–0.85 if the load-bearing bit requires a structural inference (e.g., SPDX-majority heuristic on StarCoder splits); $c_i \leq 0.50$ if the source has no explicit license or ToS governs in lieu of a license. The rule is applied source-by-source in Table 15's rightmost column,

Table 15: Per-source metadata rationale: documentation class, load-bearing bit, and confidence basis.

Source	Doc. class	Load-bearing bit	Confidence basis
CommonCrawl	Release page + ToS	RD	No explicit license; ToS silent on redistribution
C4	Dataset card (ODC-BY)	AR	Explicit ODC-BY attribution clause
RefinedWeb	Dataset card (ODC-BY)	AR	Inherited ODC-BY with processing note
Wikipedia	Wikimedia Licensing	SA	Explicit CC-BY-SA 4.0 sitewide
Gutenberg	Project Gutenberg FAQ	CU	Explicit US public-domain statement
Books3	Pushshift release page	RD	No explicit license; ambiguous provenance
StarCoder-Perm.	SPDX scan metadata	RDv	SPDX-identified Apache/MIT permissive core
StarCoder-Copy.	SPDX scan metadata	CL	SPDX-identified GPL/LGPL copyleft core
S2ORC	Allen AI dataset card	NC	CC-BY-NC per dataset card
OpenWebMath	Dataset card (CC-BY-SA)	SA	CC-BY-SA 4.0 inherited from Wikipedia math
FLAN	Google Research card	CU	Apache-2.0 explicit
ShareGPT	Terms-of-Service page	NC	ToS forbids commercial use and redistribution
OpenAssistant	HuggingFace card	SA	CC-BY-SA 4.0 on human-contributed turns
Dolly	Databricks card	AR	CC-BY-SA 3.0 with attribution

Table 16: Historical-archive inventory (Block D, 48 decisions). "Match" cells are direct agreements between archive outcome and the proxy verdict; FR marks the three false rejections (archive approved, proxy rejected) and FA marks the four false approvals (archive rejected, proxy approved). Archive span 2021-Q2 through 2025-Q4.

Deployment mode	Archive approvals	Archive rejections	Matches	Mismatch type
Research-internal	12	3	13	FR×2 (S2ORC, StarCoder-Cp.)
Hosted-API	9	5	12	FA×2 (ShareGPT; Books3)
Default redistribute	7	6	11	FA×1 (ShareGPT), FR×1 (StarCoder-Cp.)
Share-alike-strict	3	3	5	FA×1 (ShareGPT)
Totals	31	17	41/48	3 FR, 4 FA

so readers can reconstruct Table 1's confidences from the paper itself rather than from a separate evidentiary layer.