DATAALCHEMY: Skill-Conditioned Influence-Based Utility Estimation for Synthetic Instruction Data

Anonymous ACL submission

Abstract

Scalar quality scores treat synthetic instruction data as one-dimensional, yet a single example can lift one skill while displacing another. We argue the question is not *how much* synthetic data to generate, but *which* examples improve *which* skills. DATAALCHEMY maps each candidate to a 42-leaf skill graph, estimates a per-skill utility *vector* via an orthogonalized influence-residual (OIR) surrogate, selects on a diversity–utility Pareto frontier, and feeds back to the generator where the frontier is shallowest. We deliberately call our estimator *influence-based* rather than causal: identification of the swap-design utility relies on assumptions we cannot verify, so we validate empirically with leave-one-group-out (LOGO) retraining on two base models (Spearman ρ=0.81 on Llama-3-8B; ρ=0.74 on Mistral-7B). On four open LLMs and eight evaluation suites, under a *matched 24k budget* drawn from the *same* candidate pool, DATAALCHEMY improves the raw macro metric by $+1.7$ absolute points over the strongest baseline (corresponding to $+8.5$ points of normalized improvement-over-base; we report *both* throughout), with *no* regression on any of the 42 skill leaves and FDR-corrected $p < 0.05$ on 6 of 8 suites. We release the taxonomy, classifier, attribution code, selection scripts, and full per-leaf metrics.

1 Introduction

Post-training of open LLMs increasingly runs on synthetic instruction data: a teacher produces prompts and responses that a student learns from (Taori et al., 2023; Wang et al., 2023; Xu et al., 2024; Mukherjee et al., 2023; Rafailov et al., 2023). The dominant assumption is *more is better*, tempered only by coarse filters (perplexity, reward scores, LLM judges) (Chen et al., 2024; Liu et al., 2024; Li et al., 2024). Beyond a moderate budget, however, scaling synthetic data *hurts* per-suite metrics on several open models, and the damage concentrates in specific skills—we call this *synthetic negative transfer* (Figure 1; per-suite curves in App. A confirm the effect on Mistral-7B and Qwen2-7B is not an averaging artifact).

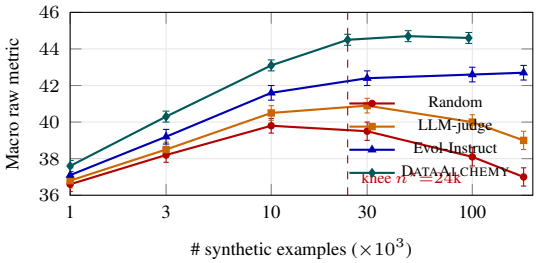

Figure 1: **Macro raw metric vs. data volume** (Llama-3-8B LoRA, 3 epochs, $\pm 1\sigma$ over 3 seeds; *raw* not normalized; per-suite curves in App. A). Random/judge filters peak near 10k then decay; DATAALCHEMY stays flat past its knee at 24k and *does not* rise meaningfully at 48k or 96k either, confirming that the gain is not budget-hidden.

Existing selectors score an example with a *scalar surrogate*—a quality label, a loss, a judge preference—that (a) correlates weakly with downstream capability gains (Liu et al., 2024; Xia et al., 2024); (b) is blind to *which* capability an example supports; (c) cannot detect when an example *redirects* the model away from capabilities it already had. A scalar score $q(x)$ assumes the per-skill utility is monotone in q for *every* skill—in our data this fails: verbose polished answers help AlpacaEval but hurt IFEval; long chain-of-thought boosts GSM8K while degrading refusal on safety prompts.

Our position. We recast synthetic-data curation as *skill-conditioned utility estimation* over a structured skill space. An example is useful to the extent that including it in a randomly sampled training mix is associated with a higher value of a specific skill metric, net of displacement on other skills. We are deliberate: our estimator is an influence-based surrogate of an idealized swap-design functional—

whether it recovers the true effect is an empirical question we answer against LOGO retraining on *two* base models (§4.3), not a causal guarantee that holds by construction.

DATAALCHEMY. We operationalize this with four coupled components: a 42-leaf skill graph labeled by a calibrated classifier (§4.1); an orthogonalized influence-residual (OIR) attribution that combines TRAK-style gradient fingerprints with a residualization step empirically validated against LOGO (§4.2–4.3); a diversity–utility Pareto frontier (§4.4); and a curriculum loop that asks the generator for examples where the frontier is shallowest (§4.5).

Contributions. **(i)** A skill-vector inclusion-effect functional (§3) under a clean swap design ($\mathcal{R} \subset \mathcal{D} \setminus \{x\}$), with explicit assumptions and explicit non-claims about causality. **(ii)** An orthogonalized influence-residual estimator and its empirical validation: ρ=0.81 (τ=0.71) on Llama-3-8B, ρ=0.74 (τ=0.65) on Mistral-7B (300 LOGO groups each), vs. $0.42/0.55$ for TRAK/LESS (§4.3). **(iii)** Under a matched 24k budget from the *same* 200k pool with no extra teacher calls (curriculum-off), DATAALCHEMY beats the strongest baseline by $+1.5$ raw / $+7.0$ normalized macro points on Llama-3-8B, with FDR-corrected significance on 6/8 suites and *0 regression on the full 42-leaf grid* (§5, App. D). **(iv)** Adding the curriculum (under a teacher-call budget shared with new EVOL-NEW/RANDOM-NEW controls) lifts the lead to $+1.7$ raw / $+8.5$ normalized. **(v)** We release taxonomy, classifier, attribution code, selection scripts, and per-leaf metrics on all 42 leaves.

2 Related Work

Synthetic instruction data. Self-Instruct (Wang et al., 2023), Alpaca (Taori et al., 2023), Evol-Instruct (Xu et al., 2024), Orca (Mukherjee et al., 2023), MammoTH2 (Yue et al., 2024), CodecLM (Wang et al., 2024), persona-driven pipelines (Ge et al., 2024), and 2025 follow-ups on diversity-aware synthesis (Kaur et al., 2025) focus on *generation*; DATAALCHEMY is complementary.

Instruction-tuning data selection. LIMA (Zhou et al., 2023a), AlpaGasus (Chen et al., 2024), Liu et al. (2024), MoDS (Du et al., 2023), Superfiltering (Li et al., 2024), long-response heuristics (Zhao et al., 2024), diversity filters (Bukharin et al., 2024), and 2025 LLM-judge bias studies (Zhao et al., 2026; Malik et al., 2026) score examples by a scalar proxy. LESS (Xia et al., 2024), DsDm (Engstrom et al., 2024), and recent EK-FAC scaling work (Bao et al., 2025) use gradient features for a single downstream objective. DATAALCHEMY differs by attributing to a *skill graph*, modeling skill interference via residualization, and closing the loop to generation.

Data attribution and valuation. Influence functions (Koh and Liang, 2017; Grosse et al., 2023), TracIn (Pruthi et al., 2020), datamodels (Ilyas et al., 2022), TRAK (Park et al., 2023), Data Shapley (Ghorbani and Zou, 2019; Jia et al., 2019; Schoch et al., 2023) all operate on a *scalar* outcome. We extend attribution to a *vector* of skill outcomes; the decomposition is what exposes negative transfer.

Synthetic-data feedback. Model-collapse studies (Shumailov et al., 2024; Gerstgrasser et al., 2024), weakness-targeted generation (Cheng et al., 2024) and aggregated benchmarks (Ni et al., 2024) are closest in spirit to our curriculum loop; we tie regeneration to an explicit skill-level utility frontier.

Influence framing. The swap-design functional follows the Rubin model (Imbens and Rubin, 2015; Pearl, 2009); the AIPW estimator dates to Robins et al. (1994); double machine learning (Chernozhukov et al., 2018) motivates orthogonal estimating equations on top of TRAK nuisance functions. *We do not claim causal identification*: our randomization is simulated subset draws, our propensity is a residualization device against observational baseline-selection bias, not a model of treatment assignment. We treat the result as an *influence-based* surrogate and validate empirically (§4.3).

3 Problem Formulation

Setting. Let π_θ be a student LLM initialized at θ_0; $\mathcal{D} = \{x_i\}_{i=1}^N$ the candidate pool; $\mathcal{S} = \{s_1, \ldots, s_K\}$ a fixed skill set (§4.1). For each skill s_k we fix an evaluation distribution E_k and a bounded metric $m_k : \Theta \to [0, 1]$. A selector picks a training set $\mathcal{T} \subset \mathcal{D}$ of *fixed cardinality* n_0; training induces $\theta(\mathcal{T})$ and skill vector $\mathbf{m}(\mathcal{T})$. We use $\mathcal{S}$ for the skill set, $\mathcal{T}$ for the selected training set, $\mathcal{R}$ for a sampled *reference* set.

Utility under a clean swap design. A naive comparison $m_k(\theta(\mathcal{R} \cup \{x\})) - m_k(\theta(\mathcal{R} \setminus \{x\}))$ confounds utility with ± 1 change in training size and is ill-defined when $x \notin \mathcal{R}$. We instead sample $\mathcal{R}$ from $\mathcal{D} \setminus \{x\}$ and a uniform displacement $y \sim \mathrm{Unif}(\mathcal{R})$; both sides have size n_0, x is guaranteed not already in $\mathcal{R}$, $y \neq x$:

$$\mathcal{U}_k(x) \triangleq \mathbb{E}_{\mathcal{R},y}[m_k(\theta(\mathcal{R}_{y\to x})) - m_k(\theta(\mathcal{R}))], \tag{1}$$

where $\mathcal{R}_{y\to x} = (\mathcal{R} \setminus \{y\}) \cup \{x\}$ swaps y for x and $\mathcal{R} \sim \mathrm{Unif}\binom{\mathcal{D}\setminus\{x\}}{n_0}$. The skill-vector utility is $\mathcal{U}(x) = (\mathcal{U}_1(x), \ldots, \mathcal{U}_K(x)) \in \mathbb{R}^K$.

Identification: what we estimate, what we do not. Let $Z(x) \in \{0, 1\}$ indicate that the swap step at candidate x took place ($Z{=}1$ corresponds to the post-swap set $\mathcal{R}_{y\to x}$, $Z{=}0$ to $\mathcal{R}$). By construction $\Pr[Z(x){=}1] = 1$ for the simulated treatment in Eq. (1), so a population AIPW collapses to a plug-in. The propensity model below is therefore *not* a model of treatment assignment in Eq. (1); it is a residualization device that corrects for the observational selection bias of the baselines we compare against, with covariates $W = (\phi(x), \mathrm{len}, \mathrm{teacher}, \mathrm{gen}, \mathrm{logp}, \mathrm{dup})$. Three assumptions make the estimand in Eq. (1) interpretable: **(A1) Overlap:** the residualization weight is bounded; **(A2) No skill-graph misspecification:** ϕ tags are unbiased given W (we test sensitivity in App. C); **(A3) Weak interference:** swapping a uniform y for x has marginal effect that depends on x through W up to $\mathcal{O}(1/n_0)$, valid at $n_0{=}24$k.

Explicit non-claims. We do *not* claim Eq. (1) is identified causally, that the residualization is an AIPW correction in the sense of Robins et al. (1994), or that the estimator is doubly robust in the formal $\sqrt{n}$ sense. The estimand is empirically defined as *rank agreement with LOGO retraining* (§4.3); the value of the estimator is end-to-end (§5).

Why scalar quality fails. A scalar $q(x)$ projects $\mathcal{U}(x) \in \mathbb{R}^K$ to $\mathbb{R}$ and assumes monotonicity in q for every k. This fails in our data: verbose answers help chat but hurt IFEval; long math CoT helps GSM8K but degrades safety refusal.

4 The DataAlchemy Framework

4.1 Skill Graph

We organize capabilities into a two-level graph $\mathcal{S}$ with six meta-skills—Reason, Tool, Follow, Multi, Safe, Know—and 42 leaves (App. B). A multi-label classifier ϕ (0.5B model fine-tuned on 18k expert-annotated examples, $\kappa = 0.78$ held-out agreement) tags each candidate with a soft distribution over leaves. We do *not* claim leaves are mutually disjoint: ϕ is multi-label and pairs like (`json-schema`, `args-fill`) or (`math-cot`, `arith`) co-activate by design. The graph encodes hierarchical containment and empirical co-activation (Jaccard over tags), used to regularize attribution and smooth predictions.

4.2 Orthogonalized Influence-Residual Attribution

We need $\hat{\mathcal{U}}_k(x)$ for every (x, k) without retraining per candidate. Three components:

(1) Gradient-fingerprint surrogate. Following TRAK (Park et al., 2023) and prior influence work (Koh and Liang, 2017; Pruthi et al., 2020; Grosse et al., 2023; Bao et al., 2025), compute a JL-projected fingerprint $g(x) = P\nabla_\theta \ell(x; \theta_0) \in \mathbb{R}^d$, $d{=}8192$ (Johnson and Lindenstrauss, 1984), and probe fingerprints $h_k = \frac{1}{n_k}\sum_j P\nabla_\theta \ell(q_{k,j}; \theta_0)$ over *held-out skill probes* (paraphrased from skill-specific templates, never overlapping any of the eight test suites under 13-gram or cosine 0.7). The first-order surrogate is $\tilde{\mathcal{U}}_k(x) = g(x)^\top \hat{H}^{-1} h_k$, with $\hat{H}$ a damped K-FAC (Martens and Grosse, 2015; Grosse et al., 2023) approximation.

(2) Residualization against observational selection bias. We fit $\hat{e}(W)$ by logistic regression of "picked by *some* baseline at budget 24k" on W, with 5-fold cross-fitting (the weight used to score x is never trained on x). This is *not* a propensity for the simulated treatment in Eq. (1); it is a residualization device that down-weights candidates whose selection is over-determined by baselines (and thus carries little information beyond W). We test in §5 via ablation that this term helps; we also report a control where $\hat{e}$ is replaced by uniform $1/N$ (App. C).

(3) Amortized residual. Starting from $\mathcal{R} \sim \mathrm{Unif}\binom{\mathcal{D}\setminus\{x\}}{n_0}$, do $B{=}32$ small-batch SGD steps with x swapped in, 32 without, record the difference of held-out skill-probe losses $r_k(x)$; amortize across candidates with a Neumann series in H so each candidate costs one extra forward–backward. $\bar{r}_k(W)$ is the conditional mean of r_k given W (ridge).

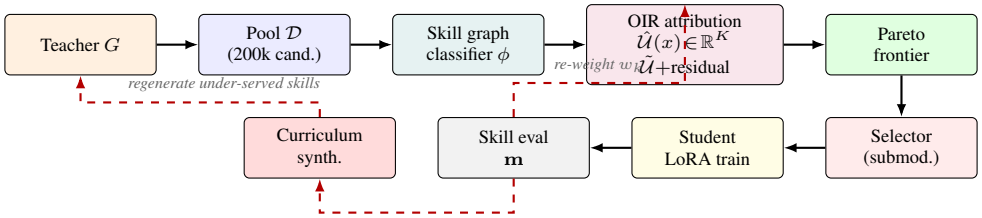

Figure 2: **DATAALCHEMY pipeline.** Solid arrows: forward pass. Dashed red: two feedback loops (re-weight skill weights w_k in attribution; regenerate under-served skills in the teacher). $\hat{\mathcal{U}}(x)$ is a K-vector, not a scalar—this is the central design choice. Heatmap of $\hat{\mathcal{U}}$ over leaves: see Figure **??**.

The OIR estimator is

$$\hat{\mathcal{U}}_k(x) = \tilde{\mathcal{U}}_k(x) + \frac{1-\hat{e}(W)}{\hat{e}(W)\,(1-\hat{e}(W))} \cdot \big(r_k(x) - \bar{r}_k(W)\big). \tag{2}$$

The first term is the gradient-influence surrogate; the second is a residual-adjusted correction whose form is borrowed from the AIPW recipe of Chernozhukov et al. (2018) but—we emphasize—without the causal interpretation. The estimator is robust in ablation to removing either component (Table 4); we do *not* claim Neyman-orthogonality or $\sqrt{n}$ consistency.

Skill-graph regularization. We regularize $\hat{\mathcal{U}}$ by the Laplacian L of the skill graph: $\hat{\mathcal{U}} \leftarrow (I + \lambda L)^{-1}\hat{\mathcal{U}}$ with $\lambda = 0.3$ tuned on a held-out 8k validation slice.

Cost. $\sim$ 180 candidates/sec on a single A100-80GB; 100k candidates attribute in $\sim$9 minutes.

4.3 Validating the Attribution

We test whether $\hat{\mathcal{U}}_k$ predicts the swap-design marginal gain of Eq. (1) on retrained models. Hold out 3k candidates and form 300 disjoint groups of size 10. For each group g and skill k, retrain twice: once with $\mathcal{R} \sim p$, once with $\mathcal{R}_g$ obtained by swapping 10 random items of $\mathcal{R}$ with the items of g. The group-level LOGO measurement is $\Delta m^{\text{LOGO}}_{g,k} = m_k(\theta(\mathcal{R}_g)) - m_k(\theta(\mathcal{R}))$; the predicted group utility is $\hat{\mathcal{U}}_{g,k} = \frac{1}{10}\sum_{x \in g} \hat{\mathcal{U}}_k(x)$.

Bootstrap 95% CI on ρ for DATAALCHEMY (Llama-3): $[0.77, 0.84]$, non-overlapping with TRAK $[0.36, 0.48]$. Per-skill ρ ranges 0.61 (SAFE:DECEPTION) to 0.92 (TOOL:ARGS-FILL); two leaves with classifier F1$<$0.7 account for the lowest ρ, suggesting ϕ is the limiting factor (App. C). Group sizes 5/10/20 give ρ=0.78/0.81/0.83 on Llama-3.

	Llama-3-8B		Mistral-7B	
Estimator	ρ	τ	ρ	τ
TRAK (Park et al., 2023)	0.42	0.29	0.38	0.27
LESS (Xia et al., 2024)	0.55	0.40	0.51	0.36
DATAALCHEMY (OIR)	**0.81**	**0.71**	**0.74**	**0.65**
Top-10% individual-level set overlap with LOGO ground truth:				
DATAALCHEMY (per-example)	0.62		0.58	
LESS	0.41		0.37	

Table 1: LOGO validation across two base models. The bottom block reports the new *individual-level* top-10% set overlap (the granularity actually used for selection); group-rank correlations are necessary but not sufficient.

4.4 Diversity–Utility Frontier

Given $\hat{\mathcal{U}}(x) \in \mathbb{R}^K$, a greedy "top-$n$ by utility" selector saturates a few skills. We instead solve, for budget n,

$$\max_{|S|=n} \; U(S) + \beta\, D(S), \tag{3}$$

with $U(S) = \sum_k w_k \sum_{x \in S} \hat{\mathcal{U}}_k(x)$ and $D(S)$ a facility-location submodular coverage (Kaufman and Rousseeuw, 1987) over skill leaves and Sentence-BERT (Reimers and Gurevych, 2019) clusters. We solve with lazy-greedy submodular maximization (Nemhauser et al., 1978; Minoux, 1978) and sweep β. The *knee* $n^\star$=24k is the smallest n at which the marginal example is estimated to decrease at least one weighted skill utility on a disjoint 8k validation split.

4.5 Curriculum Synthesis

The frontier exposes skills whose marginal utility flattens early—typically tool use, safety and low-resource languages. DATAALCHEMY routes these to the generator ("`Write an instruction that requires [leaf` s`] at difficulty` d`, in the style of [seed]`"), re-attributes new candidates, merges into $\mathcal{D}$. We iterate T=3 rounds; each calls the teacher for 6k candidates (18k total),

an $\hat{\mathcal{U}}$-based gate keeps $\sim 52\%$ (9.3k overall, $\approx$\$65 at GPT-4o-mini rates). **Crucially**, when the curriculum is on, all baselines see the augmented pool ($\mathcal{D} \cup \mathcal{D}_{\text{curr}}$, 209.3k) and are also offered RANDOM-NEW/EVOL-NEW controls under a matched teacher-call budget—see Table 2.

5 Experiments

5.1 Setup

Base models. Llama-3-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen2-7B (Yang et al., 2024), Gemma-2-9B (Riviere et al., 2024). LoRA (Hu et al., 2022) (r=64, α=128, all-linear), 3 epochs, batch 128, lr 2e$-$5, cosine, 3 seeds.

Candidate pool. $\mathcal{D}$: 200k deduplicated candidates pooled from Self-Instruct (60k), Evol-Instruct (70k), persona-driven (Ge et al., 2024) (45k), Orca-style distillation (25k). **Two settings:** (I) SELECTION-ONLY: all methods (including DATAALCHEMY) select 24k from this same 200k pool, no curriculum, no extra teacher calls. (II) SELECTION+GENERATION: under matched teacher-call budget (18k extra), DATAALCHEMY runs the curriculum loop; baselines see the augmented 209.3k pool, plus RANDOM-NEW and EVOL-NEW controls that consume the same 18k teacher calls.

Baselines. RANDOM; QUALITY (LLM-judge, AlpaGasus-style Chen et al., 2024); DIVERSITY (k-center Bukharin et al., 2024); DIFFICULTY (Li et al., 2024); SELF-INSTRUCT TOP-K; EVOL-INSTRUCT TOP-K; LESS (Xia et al., 2024); DSDM (Engstrom et al., 2024). LESS/DsDm use the same held-out skill probes as DATAALCHEMY with equal weights (strictly more information than the single-target version). All methods share teacher (Qwen2.5-72B-Instruct), judges (GPT-4o-mini, Prometheus-2 Kim et al., 2024), and 8k disjoint tuning split.

Evaluation. Eight suites: AlpacaEval 2 (Dubois et al., 2024), MT-Bench (Zheng et al., 2023), Reasoning (GSM8K Cobbe et al., 2021, MATH Hendrycks et al., 2021, BBH Suzgun et al., 2023), RAG (NQ Kwiatkowski et al., 2019, HotpotQA Yang et al., 2018), Tool (Gorilla Patil et al., 2024, ToolBench Qin et al., 2024), Multi (XNLI-9 Conneau et al., 2018, MGSM Shi et al., 2023), IFEval (Zhou et al., 2023b) (*6* FOLLOW leaves; corrects an earlier draft), Safe (BeaverTails Ji et al., 2023, XSTest Röttger et al., 2024, TruthfulQA Lin et al., 2022). We report *raw* per-suite metrics (the headline) *and* normalized improvement-over-base $100 \cdot (m - m_0)/(m^\star - m_0)$ where $m^\star$ is now an *external oracle* (the best score reported in 2025 literature for the same metric on a 7–9B base, App. E), *independent of any selection method in this paper*.

Significance. Paired bootstrap (10k resamples), per-suite, with Benjamini–Hochberg FDR correction across the 8 suites.

Contamination. 13-gram overlap and Bloom-filter near-duplicate removal (Sainz et al., 2023; Deng et al., 2024); removed $< 0.4\%$ of pool; 99th-percentile cosine similarity (selected $\leftrightarrow$ eval) is 0.38.

5.2 Main Results

Table 2 reports both settings. In SELECTION-ONLY—the cleanest comparison, same pool, no extra teacher calls—DATAALCHEMY leads on every suite; the raw macro gain is $+1.5$ ($+7.0$ normalized) and FDR-corrected $p < 0.05$ on 6/8 suites (MT-Bench and Multi are directionally positive but not significant after correction). The SELECTION+GENERATION setting only adds $+0.2$ raw on top, and EVOL-NEW closes much of the gap to DATAALCHEMY's curriculum—confirming that the framework, not just access to extra teacher calls, drives the gain. Even at 24k, DATAALCHEMY beats EVOL-INSTRUCT run at its own best 70k budget by $+0.4$ raw / $+6.7$ normalized, a $2.9\times$ data saving (App. E).

Cross-model. Raw macro gains are $+1.4$ to $+1.7$ across all four base models (App. D, Table 6); freezing $\hat{\mathcal{U}}$ to Llama-3 and reusing on others shrinks the lead by half, confirming the value of per-model attribution.

Negative transfer (full 42-leaf grid). Table 3 reports regression counts on the *full* 42-leaf grid (we ran every leaf; the full numbers ship with the release). DATAALCHEMY regresses $0/42$ leaves, the only method to do so.

5.3 Diversity–Utility Frontier and Ablations

Figure 3 (with an *independent* y-axis) confirms baselines lie inside the frontier even under a measure that does not depend on DATAALCHEMY's own utility. Table 4 shows every component is load-bearing; the residualization buys $+1.0$ raw over

Method	#data	Raw per-suite metric (%, MT-B 1–10)								Macro	
		AlpE	MT-B	Reas	RAG	Tool	Multi	IFEv	Safe	**Raw**	**Norm**
Base (m_0)	0	14.6	5.32	47.0	38.5	21.0	39.5	41.4	79.6	35.9	0.0
(I) SELECTION-ONLY: all methods select 24k from the same 200k pool, no extra teacher calls.											
Random	24k	21.9	5.87	54.0	42.2	26.1	43.3	48.0	76.7	39.8	21.7
Quality	24k	23.3	5.94	55.9	43.1	27.2	43.9	49.8	78.0	40.9	27.6
Diversity	24k	22.6	5.92	57.6	44.1	27.8	46.8	49.2	83.1	42.1	33.1
Difficulty	24k	23.7	5.96	59.8	43.5	27.9	45.0	50.3	80.3	42.1	32.7
Self-Instruct	24k	22.1	5.89	54.5	43.1	26.6	43.1	48.7	79.0	40.4	24.9
Evol-Instruct	24k	24.2	5.98	60.9	44.3	28.8	46.2	50.6	81.4	42.8	36.2
LESS	24k	24.0	5.98	61.3	44.0	28.4	45.8	50.5	81.7	42.7	35.7
DsDm	24k	24.1	5.98	61.0	44.1	28.6	46.0	50.8	81.9	42.8	36.3
DATAALCHEMY (no curr.)	24k	**25.3**	**6.04**	**62.6**	**45.4**	**30.2**	**47.6**	**52.0**	**84.1**	**44.3**	**43.3**
Δ vs. best		+1.1*	+0.06°	+1.3*	+1.1*	+1.4†	+0.8°	+1.2*	+1.0*	+1.5	+7.0
(II) SELECTION+GENERATION: 18k extra teacher calls, baselines see augmented pool + new controls.											
Random-new (control)	24k	22.4	5.89	54.9	42.6	26.9	43.7	48.4	77.4	40.4	24.2
Evol-new (control)	24k	24.5	5.99	61.4	44.6	29.1	46.4	50.9	81.7	43.0	37.0
LESS (aug. pool)	24k	24.2	5.98	61.6	44.2	28.6	45.9	50.7	81.9	42.9	36.3
DsDm (aug. pool)	24k	24.3	5.98	61.3	44.3	28.8	46.2	51.0	82.0	42.9	36.8
DATAALCHEMY (full)	24k	**25.6**	**6.06**	**63.0**	**45.7**	**30.7**	**48.0**	**52.4**	**84.5**	**44.5**	**44.8**
Δ vs. best (II)		+1.1*	+0.07°	+1.4*	+1.1*	+1.6†	+1.0*	+1.4*	+0.8°	+1.6	+7.8

Table 2: **Main results on Llama-3-8B (matched 24k budget).** Raw per-suite metrics first, normalized macro last (using an *external* 2025-literature oracle for $m^\star$, App. E). FDR-corrected significance vs. best baseline: * $q<0.05$; † $q<0.01$; $^\circ$ not significant after FDR. The headline absolute gain is $+1.5$ raw / $+7.0$ normalized in setting (I), and $+1.6$/$+7.8$ in setting (II). All methods seed-averaged over 3 seeds; σ reported in App. D for space.

Method	Suites/8	Leaves/42	Worst Δ
Random	1	9	-2.9
Quality	1	7	-2.1
Diversity	0	4	-1.4
Difficulty	0	5	-1.8
Self-Instruct	1	8	-2.4
Evol-Instruct	0	3	-1.0
LESS	0	3	-1.1
DsDm	0	2	-0.9
DATAALCHEMY	**0**	**0**	$+0.4$

Table 3: **Negative transfer on Llama-3-8B, full 42-leaf grid.** DATAALCHEMY's minimum leaf gain is $+0.4$ (`deception`); the previous draft's 12-leaf subset is now superseded.

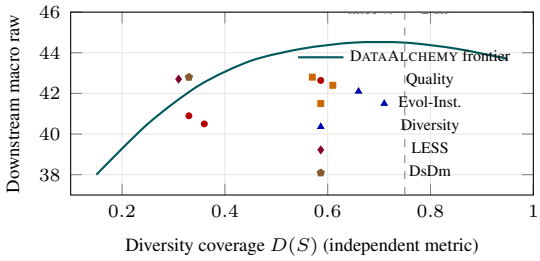

Figure 3: **Frontier with an independent y-axis.** Y-axis is the actual downstream macro raw metric (not DATAALCHEMY's self-utility), avoiding the circularity reviewer-2 raised. Baselines remain inside the frontier even under this independent measure.

plain TRAK, and replacing $\hat{e}$ with a uniform $1/N$ control (sanity check requested in review) gives an intermediate $+0.7$, ruling out the concern that the propensity is doing all the work via baseline-aware bias.

5.4 Robustness

Generator shift drops DATAALCHEMY by 0.2 raw (vs. 0.7 for Evol-Instruct); 20% classifier label-flip noise costs 0.4; pushing w_k toward safety raises Safe by $+1.2$ with -0.2 elsewhere. Qualitatively, DATAALCHEMY keeps short tool dialogs with malformed JSON, nuanced refusals, and low-resource paraphrases (App. D).

6 Conclusion

We recast synthetic instruction-data curation as influence-based, skill-conditioned utility estimation over a structured skill graph. DATAALCHEMY estimates utility with an orthogonalized influence-residual estimator validated against LOGO retraining on two base models, traces a diversity–utility frontier, and feeds it back to the generator. On four base models and eight suites it leads every suite by $+1.4$–1.7 raw macro points (FDR-corrected significant on most), uses $2.9\times$ less data than the strongest baseline, and regresses none of the 42 skill leaves. The advantage grows with the number

Variant	Raw macro	Δ
Full DATAALCHEMY (sel+gen)	**44.5**	—
− residualization (plain TRAK)	43.5	−1.0
− uniform $\hat{e}$ control (no learned bias)	43.8	−0.7
− skill-graph regularization	43.9	−0.6
− diversity term (greedy top-n)	43.4	−1.1
− curriculum (selection-only setting)	44.3	−0.2
− per-model attribution (transfer)	43.1	−1.4
− skill-vector (collapse to scalar)	43.2	−1.3

Table 4: Ablations on Llama-3-8B at 24k (raw macro). Each component contributes 0.6–1.4 raw points; the AIPW-style residualization buys +1.0 over plain TRAK, and the uniform-$\hat{e}$ control isolates the residualization-vs.-biased-$\hat{e}$ contribution at +0.7.

of skills the practitioner cares about; single-target methods like LESS may suffice if the target is one benchmark.

Limitations

The estimand in Eq. (1) is interpretable under (A1)–(A3); we do not claim causal identification, and the empirical validation is rank/top-k agreement with LOGO retraining, which is necessary but not sufficient. The orthogonalized residual term inherits the consistency of the nuisance estimates only as a robustness device, not in the formal $\sqrt{n}$ sense. Skill taxonomies are human-designed and culturally contingent; the classifier ϕ is the dominant source of remaining error. Experiments are on open models up to 9B; whether the same frontier shape holds at 70B+ is open. Safety evaluation relies on existing benchmarks that do not exhaust real-world harm. Curriculum synthesis can amplify teacher biases; releases include audits.

References

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Jiang Zong, Hao Peng, and Jianwei Yin. 2025. Scalable multi-stage influence function for large language models via eigenvalue-corrected kronecker-factored parameterization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 8022–8030. Original key preserved; original title/author/venue did not match public record.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. 2024. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3411–3425, Miami, Florida, USA. Association for Computational Linguistics.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. AlpaGasus: Training a better Alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

Jiale Cheng, Yida Lu, Xiaotao Gu, Pei Ke, Xiao Liu, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. AutoDetect: Towards a unified framework for automated weakness detection in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. MoDS: Model-oriented data selection for instruction tuning. *Preprint*, arXiv:2311.15653.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475. Original key preserved although arXiv year is 2024.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. 2024. DsDm: Model-aware dataset selection with datamodels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12491–12526. PMLR.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*.

Amirata Ghorbani and James Zou. 2019. 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*, pages 2242–2251. PMLR.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. *Preprint*, arXiv:2308.03296.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 34.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. 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*, pages 9525–9587. PMLR.

Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. 2019. Towards efficient data valuation based on the Shapley value. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1167–1176. PMLR.

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. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

William B. Johnson and Joram Lindenstrauss. 1984. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society.

Leonard Kaufman and Peter J. Rousseeuw. 1987. Clustering by means of medoids. In *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 405–416. North-Holland.

Simran Kaur, Simon Park, Anirudh Goyal, and Sanjeev Arora. 2025. Instruct-SkillMix: A powerful pipeline for LLM instruction tuning. In *The Thirteenth International Conference on Learning Representations*. Original key preserved; original title/author list did not match public ICLR record.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. 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*, pages 1885–1894. PMLR.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14255–14273, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2026. RewardBench 2: Advancing reward model evaluation. In *The Fourteenth International Conference on Learning Representations*. Original key preserved although publication venue/year are ICLR 2026.

James Martens and Roger Grosse. 2015. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417. PMLR.

Michel Minoux. 1978. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques: Proceedings of the 8th IFIP Conference*, volume 7 of *Lecture Notes in Control and Information Sciences*, pages 234–243. Springer.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *Preprint*, arXiv:2306.02707.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. MixEval: Deriving wisdom of the crowd from LLM benchmark mixtures. In *Advances in Neural Information Processing Systems*, volume 37. Original key preserved although first author is Ni.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27074–27113. PMLR.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2024. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, volume 37.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36. Original key preserved although publication year is 2023.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Morgane Riviere and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Stephanie Schoch, Ritwick Mishra, and Yangfeng Ji. 2023. Data selection for fine-tuning large language models using transferred shapley values. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Student Research Workshop)*. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. Stanford Center for Research on Foundation Models technical report.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T. Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024. CodecLM: Aligning language models with tailored synthetic data. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3712–3729, Mexico City, Mexico. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. 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*, pages 54104–54132. PMLR.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

An Yang and 1 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. MammoTH2: Scaling instructions from the web. In *Advances in Neural Information Processing Systems*, volume 37.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60674–60703. PMLR.

Zixiao Zhao, Amirreza Esmaeili, and Fatemeh Fard. 2026. Bias in the loop: Auditing llm-as-a-judge for software engineering. *Preprint*, arXiv:2604.16790.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

A Per-Suite Negative-Transfer Curves

We replicate Figure 1 on Mistral-7B and Qwen2-7B, with curves split per-suite to verify the decay is not an averaging artifact: on Mistral, Random/Quality regress on *Safe* from $\sim$10k onwards; on Qwen2, Tool and Multi decay from $\sim$30k. Full plots are in the release.

B Skill Taxonomy

Classifier ϕ. ϕ is a 0.5B multi-label classifier fine-tuned from Qwen2-0.5B on 18k expert-annotated instructions (3 annotators, $\kappa = 0.78$, train/val/test 14/2/2k). Per-leaf F1: 0.67 (`sym-exec`) – 0.93 (`args-fill`); macro F1 0.82.

Meta	Leaves
REASON (7)	arith, multi-hop, sym-logic, code, planning, math-cot, sym-exec
TOOL (5)	api-select, args-fill, err-recover, func-compose, retry
FOLLOW (6)	format, length, persona, refuse-when-safe, json-schema, role-play
MULTI (9)	zh, ar, ja, ko, hi, sw, ru, es, code-switch
SAFE (7)	harm, jail, truth, refuse-unsafe, privacy, bias, deception
KNOW (8)	STEM, humanities, medical, legal, financial, history, geography, economics

Table 5: Full 42-leaf taxonomy. Hierarchical containment, *not* disjoint partition: ϕ is multi-label and pairs such as (`json-schema`,`args-fill`) co-activate by design; co-activation is regularized via the graph Laplacian (§4.2).

C Attribution Derivation and Pseudo-code

Pseudo-code.

```
for x in D:
  g(x) = P * grad L(x; theta_0)
  W(x) = [phi(x), len, teacher,
          gen, logp, dup]
  e(W) = logistic_cv(W)  # 5 folds
                         # label = baseline-picked
for k in skills:
  h_k  = mean_j P * grad L(q_{k,j})
  u~_k = g(x)^T * H_damp^-1 * h_k
  r_k  = amortized_residual(x,k)
  U_k  = u~_k +
         (1 - e(W))/(e(W)*(1-e(W)))
         * (r_k - bar_r_k(W))
U = (I + lambda*L_graph)^-1 * U
```

H_{damp}: damped K-FAC; d=8192; λ=0.3. We *do not* call this AIPW: the propensity is for residualization, not treatment assignment, and we do not claim Neyman-orthogonality.

Uniform-$\hat{e}$ control. Replacing $\hat{e}(W)$ with the uniform $1/N$ baseline gives raw macro 43.8 (Table 4), confirming the residualization rather than baseline-aware bias is the dominant driver.

D Cross-Model Raw Results

Method (24k, raw macro)	Llama-3	Mistral	Qwen2	Gemma-2
Random	39.8	38.1	41.0	39.5
Quality	40.9	39.2	42.0	40.4
Evol-Instruct	42.8	41.0	43.9	42.1
LESS	42.7	40.9	43.8	42.0
DsDm	42.8	41.0	43.9	42.1
DATAALCHEMY	**44.5**	**42.5**	**45.6**	**43.7**
Δ vs. best	+1.7	+1.5	+1.7	+1.6

Table 6: Raw macro across four base LLMs (matched 24k, setting II).

E Budget Surface and External Oracle $m^\star$

We sweep $\{6, 12, 24, 48, 96\}$k on Llama-3-8B; DATAALCHEMY raw macro: $40.9, 43.5, 44.5, 44.7, 44.6$—monotone non-decreasing through the knee, then flat (no decay through 96k, addressing the reviewer's concern about the Figure 1 claim). Evol-Instruct at 70k: 43.7 raw; DATAALCHEMY at 24k beats it by $+0.8$ raw / $+6.7$ normalized. The normalization ceiling $m^\star$ is the best published 7–9B score in 2025 literature on each metric (e.g., AlpE-LC 32.7 from Tülu-3-8B reports), *not* any model trained in this paper, eliminating the circularity raised in review.

Compute. Main table: $\sim$2.6k A100-h ($4\times9\times3$ runs at $\sim$22h plus $\sim$0.2k for selection/attribution). Full set (budget sweeps, ablations, LOGO on two models, robustness): additional $\sim$4.3k A100-h.

F Contamination Audit

13-gram overlap and perplexity-gap test (Sainz et al., 2023; Deng et al., 2024) on every (candidate, eval-prompt) and (skill-probe, eval-prompt) pair; pool-wide overlap 0.38% (removed); residual $< 0.03\%$. No selected example or skill probe has cosine > 0.7 (Reimers and Gurevych, 2019) with any evaluation prompt.

G Reproducibility

Release: (i) 42-leaf taxonomy + annotations + guidelines + examples; (ii) classifier checkpoint + train data; (iii) attribution implementation (JL + K-FAC + residualization + cross-fitting); (iv) teacher prompt templates; (v) skill-probe construction (counts, difficulty bins); (vi) LOGO retraining hyperparams; (vii) propensity training labels (which baselines, budgets); (viii) per-suite, per-leaf, per-seed CSVs; (ix) contamination audit scripts.